The genome of the social amoeba *Dictyostelium discoideum*
Supplementary information

L. Eichinger[1, †], J.A. Pachebat[2,1, †], G. Glöckner[3, †], M.-A. Rajandream[4, †], R. Sucgang[5, †], M. Berriman[4], J. Song[5], R. Olsen[6], K. Szafranski[3], Q. Xu[7, 8], B. Tunggal[1], S. Kummerfeld[2], M. Madera[2], B. A. Konfortov[2], F. Rivero [1], A. T. Bankier[2], R. Lehmann[3], N. Hamlin[4], R. Davies[4], P. Gaudet[9], P. Fey[9] , K. Pilcher[9], G. Chen[5], D. Saunders[4], E. Sodergren[7,10], P. Davis[4], A. Kerhornou[4], X. Nie[5], N. Hall[4, a], C. Anjard[6], L. Hemphill[5], N. Bason[4], P. Farbrother[1], B. Desany[5], E. Just[9], T. Morio[11], R. Rost[12], C. Churcher[4], J. Cooper[4], S. Haydock[13], N. van Driessche[7], A. Cronin[4], I. Goodhead[4], D. Muzny[10], T. Mourier[4], A. Pain[4], M. Lu[5], D. Harper[4], R. Lindsay[5], H. Hauser[4], K. James[4], M. Quiles[10], M. Madan Babu[2], T. Saito[14], C. Buchrieser[15], A. Wardroper[16, 2], M. Felder[3], M. Thangavelu[17], D. Johnson[4], A. Knights[4], H. Loulseged[10], K. Mungall[4], K. Oliver[4], C. Price[4], M.A. Quail[4], H. Urushihara[11], J. Hernandez[10], E. Rabbinowitsch[4], D. Steffen[10], M. Sanders[4], J. Ma[10], Y. Kohara[18], S. Sharp[4], M. Simmonds[4], S. Spiegler[4], A. Tivey[4], S. Sugano[19], B. White[4], D. Walker[4], J. Woodward[4], T. Winckler[20], Y. Tanaka[11], G. Shaulsky[7, 8], M. Schleicher[12], G. Weinstock[7, 10], A. Rosenthal[3], E.C. Cox[21], R. L. Chisholm[9], R. Gibbs[7, 10], W. F. Loomis[6], M. Platzer[3, ‡], R. R. Kay[2, ‡], J. Williams[22, ‡], P. H. Dear[2, ‡,§], A. A. Noegel[1, ‡], B. Barrell[4, ‡] and A. Kuspa[5, 7, ‡]

[1]Center for Biochemistry and Center for Molecular Medicine Cologne, University of Cologne, Joseph-Stelzmann-Str. 52, 50931 Cologne, Germany
[2]Laboratory of Molecular Biology, MRC Centre, Cambridge CB2 2QH, UK
[3]Genome Analysis, Institute for Molecular Biotechnology, Beutenbergstr. 11, D-07745 Jena, Germany
[4]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK
[5]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX77030, USA
[6]Section of Cell and Developmental Biology, Division of Biology, University of California, San Diego, La Jolla, CA 92093, USA
[7]Dept. of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX 77030,USA
[8]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston TX 77030, USA
[9]dictyBase, Center for Genetic Medicine, Northwestern University, 303 E Chicago Ave,Chicago, IL 60611, USA
[10]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA
[11]Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8572, Japan
[12]Adolf-Butenandt-Institute/Cell Biology, Ludwig-Maximilians-University, 80336 Munich, Germany
[13]Biochemistry Department, University of Cambridge, Cambridge CB2 1QW, UK.
[14]Division of Biological Sciences, Graduate School of Science, Hokkaido University, Sapporo 060-0810 Japan
[15]Unité de Genomique des Microorganismes Pathogenes, Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris Cedex 15, France.
[16]Department of Biology, University of York, York YO10 5YW, UK.
[17]MRC Cancer Cell Unit, Hutchison/MRC Research Centre, Hills Road, Cambridge CB2 2XZ, UK.
[18]Centre for Genetic Resource Information, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
[19]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Minato, Tokyo 108-8639, Japan
[20]Institut für Pharmazeutische Biologie, Universität Frankfurt (Biozentrum), Frankfurt am Main, 60439, Germany
[21]Department of Molecular Biology, Princeton University, Princeton, NJ08544-1003, USA
[22]School of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK

[a]Present address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville MD 20850, USA

[†]These authors contributed equally.
[‡]Co-senior authors
[§]Corresponding author.
Telephone: [0044] 1223 402190
Fax: [0044] 1223 412178
Email: phd@mrc-lmb.cam.ac.uk

TABLE OF CONTENTS

"FILE" *indicates that the figure or table is too large to be included in this document, and may be downloaded as a separate file. In some cases, portions of the figure or table are reproduced here for illustration.*

Congruence of the physical map with the assembled sequence

A comparison of the HAPPY map with the sequence map of the *Dicyostelium* genome is shown in Figure SI 1. The markers used to construct the map were sequenced-tagged sites (STS's) identified by hemi-nested PCR. The primer triplets used to score each STS marker are listed in order of map location in Table SI 1, which must be downloaded as a separate file and which contains explanatory notes in addition to the data.

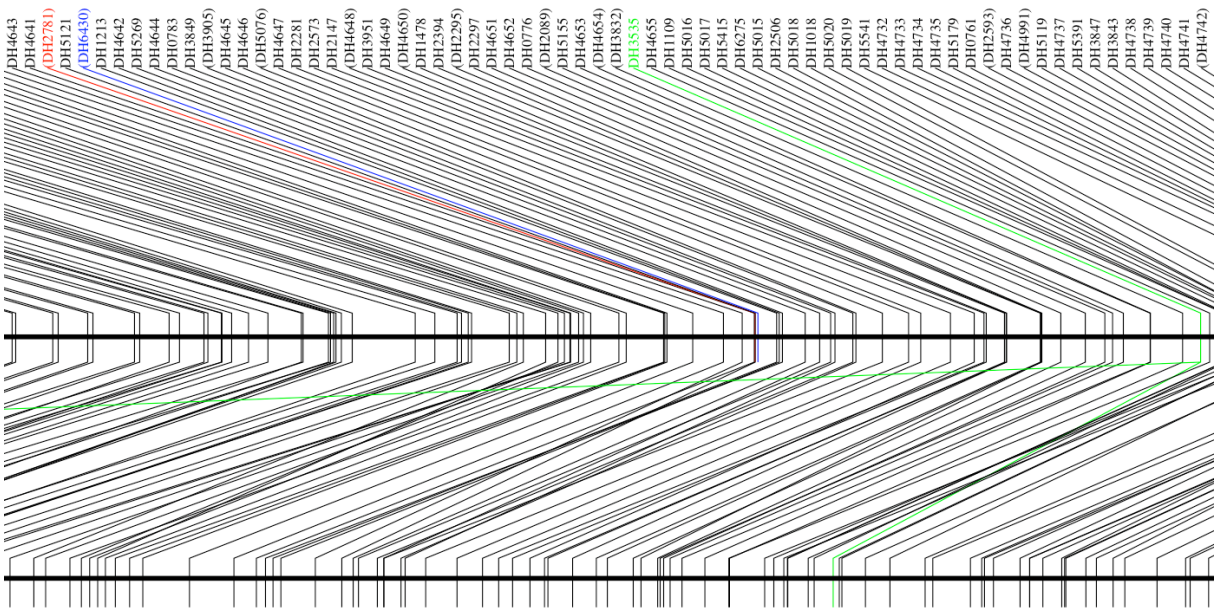


Figure SI 1. Comparison of HAPPY maps with the assembled sequence. *(Only a portion of chromosome 1 from the high-resolution figure is shown here for explanatory purposes. Due the size of the computer file, the complete figure must be downloaded separately.) For each of the six chromosomal assemblies (numbered at left), the mapped HAPPY markers are named at top. Lines indicate the position of each marker on the HAPPY map (upper horizontal line), and the location at which the corresponding sequence is found in the chromosomal assembly (lower horizontal line). Markers whose sequence cannot be found in the chromosomal assemblies (missing markers) are indicated in blue; markers whose sequence occurs twice on the chromsome (double-copy markers) are indicated in green (both positions of such markers in the assembly are shown); markers whose sequence lies on a different chromosome (wrongly-mapped markers) are indicated in red. Second-rate markers (mapped with lower confidence) are indicated by brackets around their names. The red bar beneath Chromosome 2 indicates the large inverted duplication on this chromosome; markers within this region are represented by a single map location but occur twice within the duplicated region. The distance scale is indicated (right).*

Table SI 1. Summary of STS markers used in constructing HAPPY map
(only a small part of the table is shown for illustration; the complete table may downloaded as an Excel file)

Marker name	Fwd External primer	Fwd Internal primer	Reverse primer	Size (bp)	Data quality	Chromo-some (map)	Position/kb (map)	Chromo-some (seq)	Position/kb (seq)	2nd chromo-some (seq)	Second position/kb (seq)
….	….	….	….	….	….	….	….	….	…..…	….	
DH5236	CCATCAGTTCCATCAGTTGG	CAAATTTCAATCAATGTTGGTGG	GGTTGTGATGATAATGATCC	134	1	3	2884.63	3	2740.54		
DH4966	CGTTCTTCAAGTGATTGAGG	CTTGTAGATTCATCCATTAAAGC	GGTTAAATCAATTATCACCAGG	97	1	3	2886.41	3	2732.81	3	2735
DH6029	GTTAGAGTTAATTGGTCAATTGG	GTTAGTTATCATTCATCAATACC	GTTGAATAATCTTTATCGGTTGG	86	1	3	2886.61	3	2757.55		
DH5709	CTTTCTCCATCTTTGAGTTGG	GTGGATGAGATAAGTCTTGG	GGTTGGTTAGTTCATCCAGC	101	1	3	2895.93	3	2745.38		
DH5523	CCAGAGTTTACATTACCAGG	GAATCTGTTAAACTTCTTTGGG	CCATCAATAGTCATTCTTTCC	206	1	3	2896.82	3	2754.57		
DH3027	CAATCACCTCTACTACCACC	GATGACAATGATGAAGATGG	CTCTGTTGATTGTTTACCTGC	93	1	3	2902.46	3	2752.99		
…..	…..	…..	…..	…..	…	…	…..	…..	…..	….	….

For each marker (listed in order of map location in the genome), the sequences of the hemi-nested PCR primers (forward-external, forward-internal, and reverse) are given, followed by the expected size of the internal PCR product (between the forward-internal and reverse primers). Data quality refers to the quality of the marker typing on the mapping panel; 2nd-rate markers are mapped with less confidence. Chromosome (map) and Position/kb (map) give the expected location of the marker (chromosome number and position on chromosome in kb, respectively) based on the map data. Chromosome (seq) and Position/kb (seq) give the actual location (chromosome number and position, respectively) at which the marker is found in the genome assembly, version 2. (A chromosome number of 0 indicates that the marker has not been found in the genome assembly; 7 indicates that it lies on a floating contig.) Where a marker has been found at two locations in the genome assembly, the chromosome and position of the second occurrence is given in the final two columns.

Simple sequence repeats

Simple sequence repeats were classified (monomer, dinucleotide, trinucleotide, etc) and the frequency of their occurrence in the genomes is plotted in Figure SI 2.

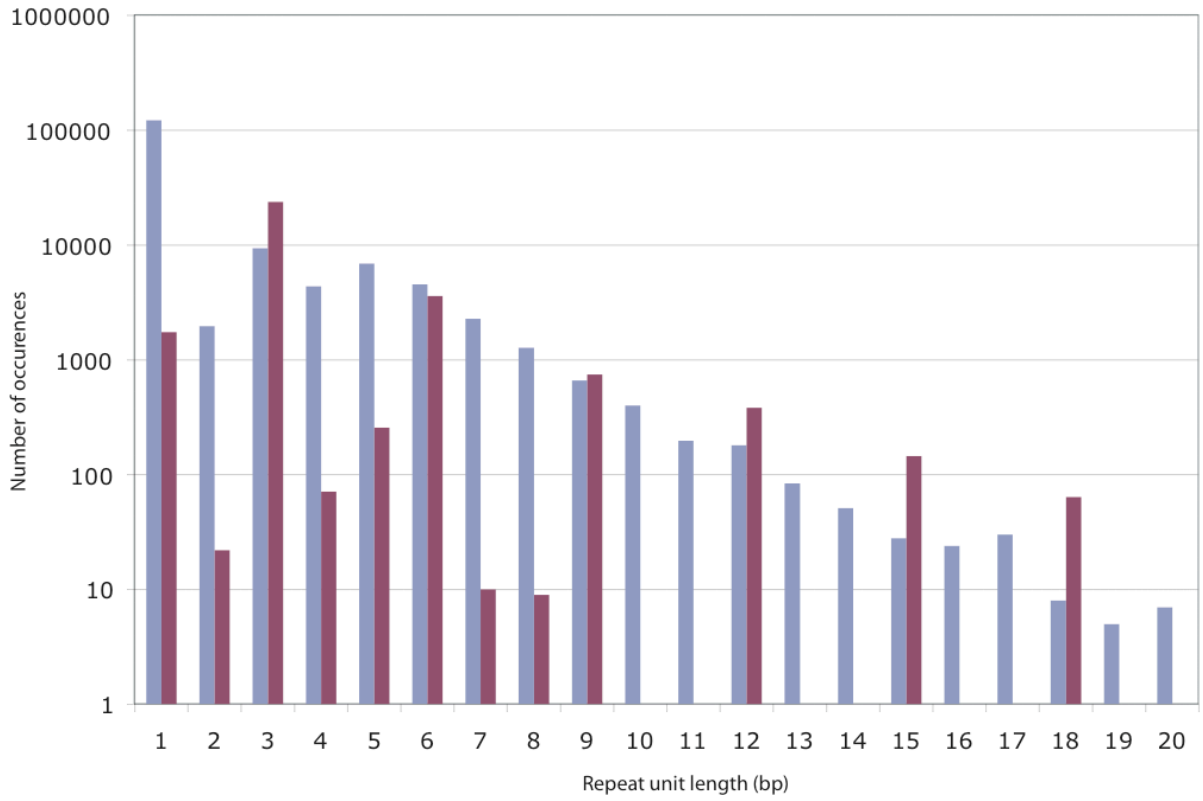


Figure SI 2. Distribution of simple-sequence repeats*. The graph shows the total number of tracts of homopolymers and tandemly repeated motifs up to 20bp, in non-coding (grey) and coding (red) sequence. The minimum number of repeats of the unit motif was 10 (homopolymers), 7 (dinucleotides), 5 (trinucleotides), 4 (tetranucleotides), or 3 (pentanucleotides and longer motifs).*

Selenocysteine insertion machinery

Systematic incorporation of selenium into proteins is facilitated by a tRNA on which the initial serine is converted to selenocysteine (hence designated $tRNA^{[Ser]Sec}$). In the presence of a selenocysteine insertion sequence (SECIS) in the mRNA, the $tRNA^{[Ser]Sec}$ recognises UGA which otherwise acts as a stop codon. Eukaryotic proteins with incorporated selenocysteines have been found in animals, plants and protozoans[1-5].

We looked for evidence of selenocysteine incorporation into *Dictyostelium* proteins and have identified a possible $tRNA^{[Ser]Sec}$ in *D. discoideum* genome and two putative selenoproteins with characteristic SECIS elements at the 3'-end of these genes (Figure SI 3). A putative selenocysteine tRNA was identified by tRNAscan-SE prediction using the covariance model analysis only for maximum sensitivity[6]. Potential SECIS elements downstream of DDB0218378 & DDB0202474 were found through an initial assessment of the folding potential of subsequences embedded in TGAN-NGAN motifs.

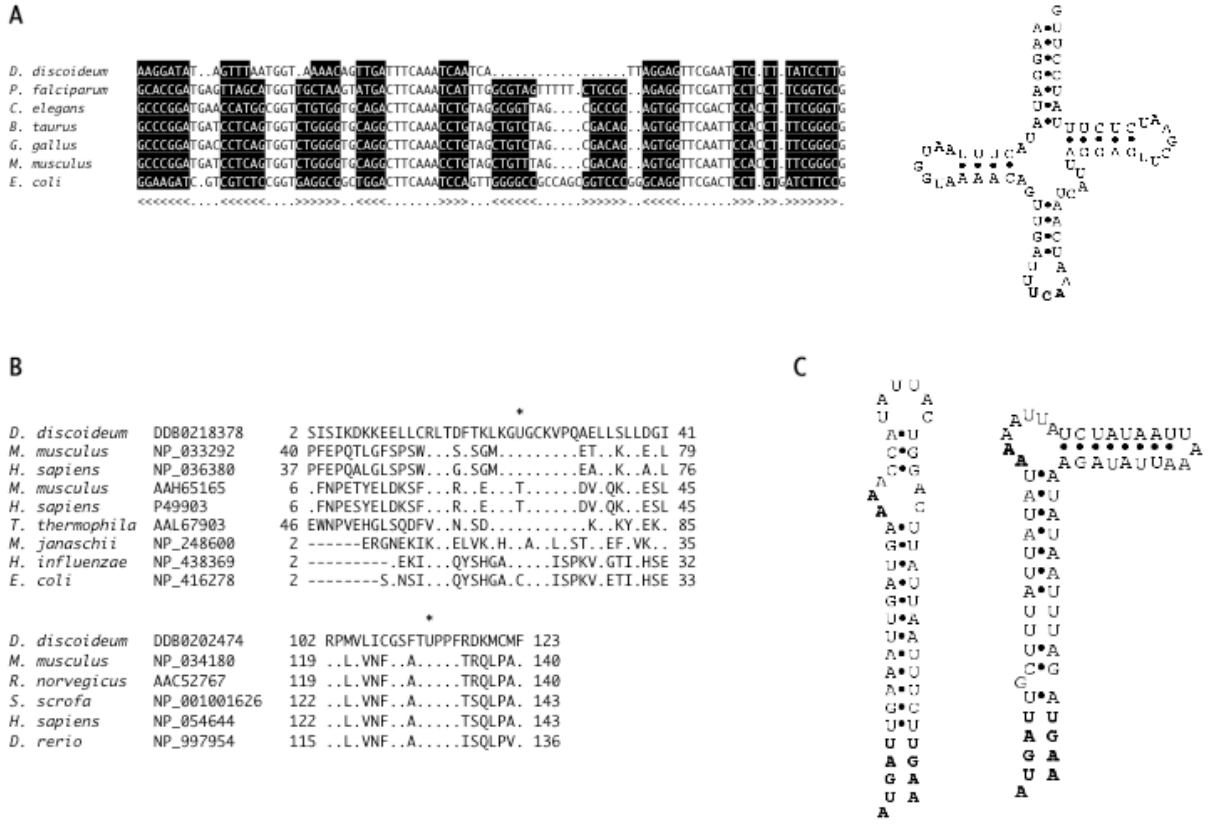


Figure SI 3. Selenocysteine insertion machinery.
A) *Alignment of selenocysteine (Sec) tRNAs (left). The predicted secondary structure consensus is shown below as nested parentheses and outlined for the putative* D. discoideum *Sec-tRNA (right). Highlighted residues in the alignment are able to base pair in accordance with the predicted structure. The D. discoideum Sec-tRNA (BC4V2_0; position 2480622-2480709) was predicted by tRNAscan-SE[6] (maximum sensitivity and lowered cut-off score).* ***B)*** *Alignment of subsequences of selenoproteins (coordinates shown around sequence). Only residues differing from the D. discoideum sequence are shown. Selenocysteine codon (Amino acid code: U) positions marked with asterisks. Top: DDB0218378 aligned with selenophosphate synthetases. Two copies of selenophosphate synthetase exist in human and mouse of which only one copy displays the selenocysteine codon. As the selenium-containing enzyme is found in all three domains of life, this most probably reflects the ancestral state[7]. Translated BLAST searches with DDB0218378 against the genome suggested that D. discoideum does not encode more than one copy. A partial sequence from DDB0218378 has previously been published (SELD_DICDI. Acc. no.: Q94497) in which the selenocysteine codon (TGA) was reported as GGA, hence translated as glycine. Bottom: DDB0202474 aligned with vertebrate type II iodothyronine deiodinases.* ***C)*** *Predicted structures of putative selenocysteine insertion sequence (SECIS) elements situated downstream of DDB0218378 (left) and DDB0202474 (right; structure predicted by SECISearch). Bold letters denote the conventional eukaryotic non-Watson-Crick quartet (UGAN-NGAN) preceded by adenosine and the invariant AA pair[8].*

Sequences with the ability to form a hairpin structure displaying the invariant AA motif in an internal bulge were selected. The putative SECIS element downstream of DDB0218378 is supported by homology to the published putative SECIS elements as well as a substitution

pattern strongly indicative of a conserved secondary structure (not shown)[9]. The DDB0202474 putative SECIS element is supported SECISearch 2.19 (http://genome.unl.edu/SECISearch.html[10]), although with very low COVE score. The putative SECIS elements identified in *D. discoideum* genome conform to the Form 2 SECIS element structure (reviewed in ref. 8). To our knowledge, this is the first report of identification of selenocysteine insertion machinery in an amoebozoan at the sequence level.

Centromere-like behaviour of DIRS element clusters

DIRS elements are restricted almost exlusively to one end of each *Dictyostelium* chromosome, and these chromosomal ends appear to cluster both in cells with condensed chromosomes characeristic of mitosis, and in interphase (Figure SI 4), suggesting that DIRS elements perform a centromeric function.

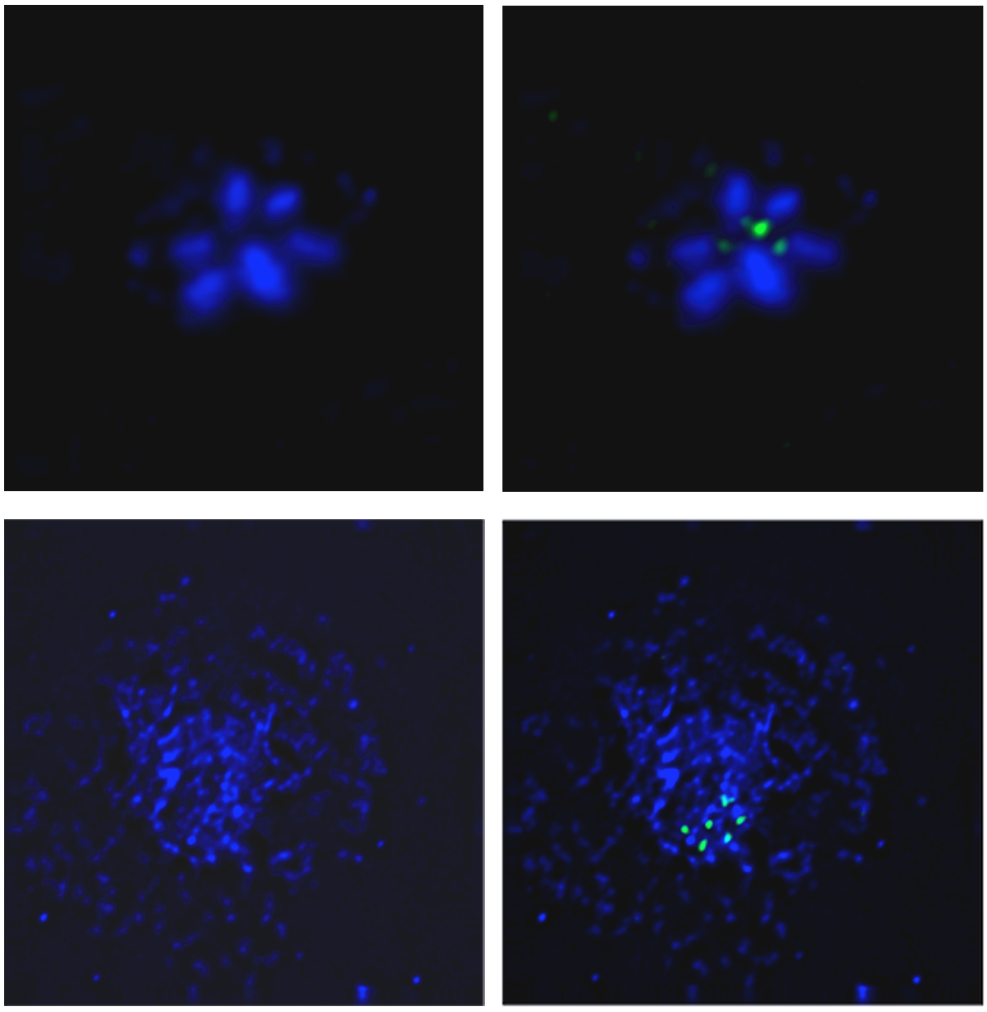


Figure SI 4. Localization of DIRS elements by *in situ* hybridization.
The top panels show a representative *Dictyostelium* cell fixed during mitosis; DNA is labelled blue (left), and hybridized with probes against DIRS elements (green; merged image at right). The lower panels show a representative cell during interphase (left: DNA labelled blue; right: merged image with DIRS element probes in green).

rDNA palindrome sequence elements at the ends of chromosomes

By sorting through the contigs that contained complex repetitive elements we found 13 high-quality contigs that ended with a portion of the rDNA palindrome. Two of these mapped to well established internal chromosomal segments, but the other 11 appear to lie at chromosome ends (Figure SI 5). One of the telomeric contigs has twice the depth of coverage as the others so we assume there may be two identical copies of this contig at distinct chromosome ends.

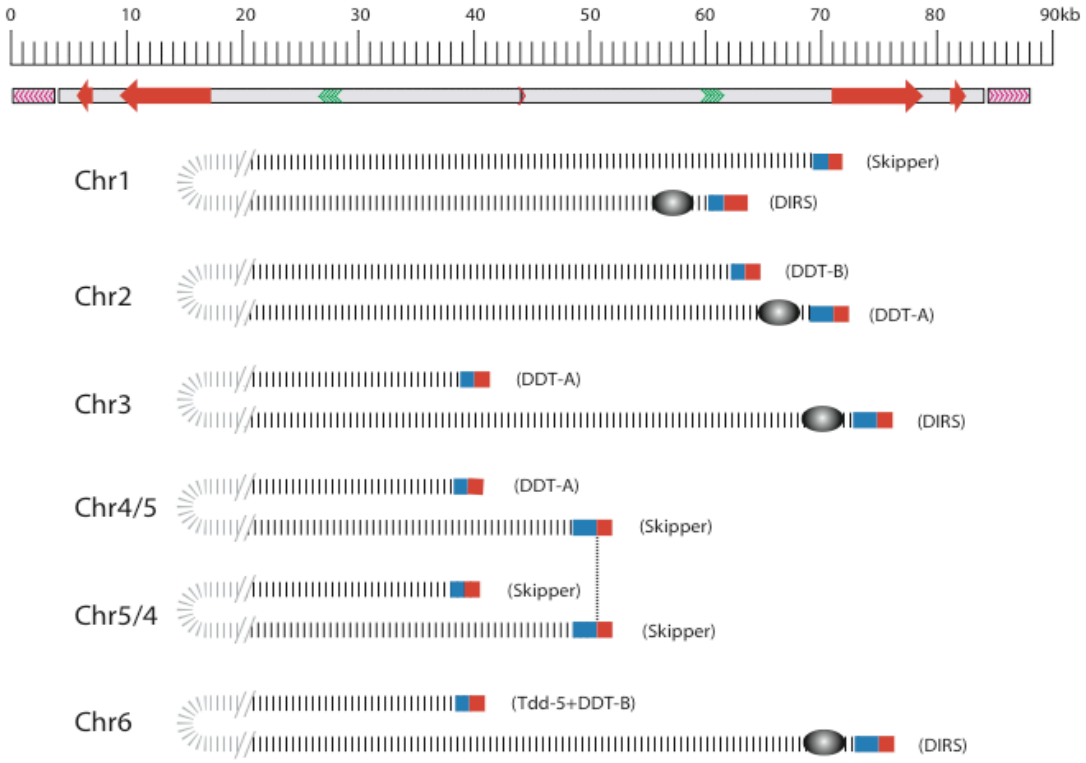


Figure SI 5 Relationship between putative telomeric sequences and the extrachromosomal rDNA element. *The complete rDNA element is shown at top (red solid arrows: rDNA genes; green chevrons: GC-rich region; pink chevrons: telomeric repeats; red chevron: central region of assymetry; a sequence gap near the tip of each arm is believed to comprise short repeats). Below are shown the junction contigs identified from the sequence, each of which comprises a transposable element (blue boxes; element type named to right) fused to a segment of the rDNA sequence (red boxes, aligned vertically with the corresponding part of the rDNA palindrome). The inferred arrangement of these junction contigs on the six chromosomes is indicated, with the DIRS-proximal end of each chromosome indicated (grey ovoid), where this could be inferred based on the repeat types present in the junction contigs (the junctions on chromosomes 4 and 5 carry only Skipper and DDT-A elements, which may be present at either end of the chromosome, and hence cannot be oriented in this way). The vertical dotted line connects two junction contigs of identical sequence.*

Proposed mechanism for the creation of the chromosome 2 duplication

Figure SI 6 shows the proposed sequence of events leading to the inverted duplication seen on chromosome 2 of *D. discoideum* AX4.

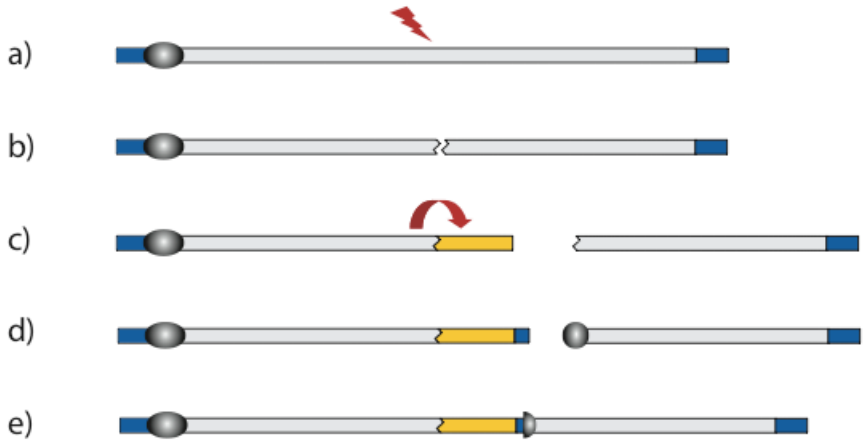


Figure SI 6. Proposed 'breakage-fusion-bridge' cycle on chromosome 2.
The original chromosome (a) is broken (b), leaving an unstable end which induces an inverted duplication (c) of the adjoining 700kb region. The new chromosomal fragments are partially stabilised by acquisition of new telomeric (blue) and centromeric (grey circle) sequences (d) but ultimately refuse to produce an extended chromosome (e) carrying residual internal telomeric and centromeric sequences.

Codon usage and tRNAs

Table SI 2. Codon usage and predicted tRNAs.

Phe	TTT 33.9 (0)	**Ser**	TCT 15.4 (9)	**Tyr**	TAT 30.4 (0)	**Cys**	TGT 12.8 (0)
	TTC 13.6 (13)		TCC 3.9 (0)		TAC 5.2 (12)		TGC 1.5 (8)
Leu	TTA 56.8 (18)		TCA 50.2 (15)	**Stp**	TAA 0.0 (0)	**Stp**	TGA 0.0 (0)
	TTG 10.6 (4)		TCG 2.3 (1)		TAG 0.0 (0)	**Trp**	TGG 7.3 (7)
Leu	CTT 9.5 (11)	**Pro**	CCT 5.9 (1)	**His**	CAT 15.0 (0)	**Arg**	CGT 5.7 (6)
	CTC 3.3 (0)		CCC 1.2 (0)		CAC 2.7 (9)		CGC 0.1 (0)
	CTA 5.2 (3)		CCA 31.9 (15)	**Gln**	CAA 48.6 (13)		CGA 0.5 (1)
	CTG 0.4 (1)		CCG 0.5 (0)		CAG 1.9 (1)		CGG 0.1 (0)
Ile	ATT 51.9 (17)	**Thr**	ACT 20.7 (16)	**Asn**	AAT 101.9 (0)	**Ser**	AGT 22.4 (0)
	ATC 11.0 (0)		ACC 7.7 (0)		AAC 11.8 (18)		AGC 2.5 (12)
	ATA 21.8 (4)		ACA 30.3 (6)	**Lys**	AAA 65.1 (22)	**Arg**	AGA 19.9 (10)
Met	ATG 15.7 (14)		ACG 1.0 (1)		AAG 11.5 (10)		AGG 1.4 (1)
Val	GTT 23.7 (22)	**Ala**	GCT 9.9 (15)	**Asp**	GAT 47.4 (0)	**Gly**	GGT 32.3 (0)
	GTC 3.2 (0)		GCC 3.2 (0)		GAC 4.5 (22)		GGC 2.1 (18)
	GTA 13.3 (6)		GCA 16.4 (0)	**Glu**	GAA 49.1 (19)		GGA 9.0 (5)
	GTG 2.3 (1)		GCG 0.6 (0)		GAG 8.8 (3)		GGG 1.0 (0)

For each codon (encoded amino acid named at left; Stp=stop codon), the frequency of its occurrence per 1000 codons in all *D. discoideum* predicted exons is given; the number in brackets indicates the number of predicted tRNA genes with the complementary anticodon (i.e. capable of decoding the corresponding codon without wobble). Heavy lines bound groups of synonymous codons differing only in their third base; within these groups, the most common codon and most abundant tRNA are highlighted in red.

Analysis of gene duplications

Raw data for the analysis of inferred gene duplication in the *Dictyostelium* genome is given in. Table SI 3, which must be downloaded as a separate file. The 13,498 predicted proteins were run all-against-all BLASTP with -e 1e-5 and -F T. The result was fed into TribeMCL with -e 1e-40 (see Methods). Sheet 1 of the spreadsheet lists all the members of all families that had at least 3 members. Families were given a number beginning with 0. There are a total of 351 such families. Sheet 2 lists pairs of genes that are found in the same family and which lie on the same chromosome. The first three columns list the family number (as for sheet 1) and the identifiers of the members of the pair. The fourth and fifth columns give, respectively, the physical distance between the two genes in basepairs and an inferred phylogenetic distance from the program protdist (part of the PHYLIP package).

Distribution of amino acid homopolymers in the *Dictyostelium* proteome

Table SI 4. Amino acid composition and minimum non-random homopolymer length.

AA	Composition	p<0.05	p<0.01	p<0.001
A	0.030	6	6	7
C	0.014	5	5	6
D	0.052	7	7	8
E	0.058	7	8	8
F	0.047	7	7	8
G	0.044	6	7	8
H	0.018	5	6	6
I	0.085	8	9	10
K	0.077	8	8	9
L	0.086	8	9	10
M	0.016	5	5	6
N	0.113	9	10	11
P	0.039	6	7	7
Q	0.050	7	7	8
R	0.028	6	6	7
S	0.097	8	9	10
T	0.060	7	8	8
V	0.042	6	7	8
W	0.007	4	5	5
Y	0.036	6	7	7

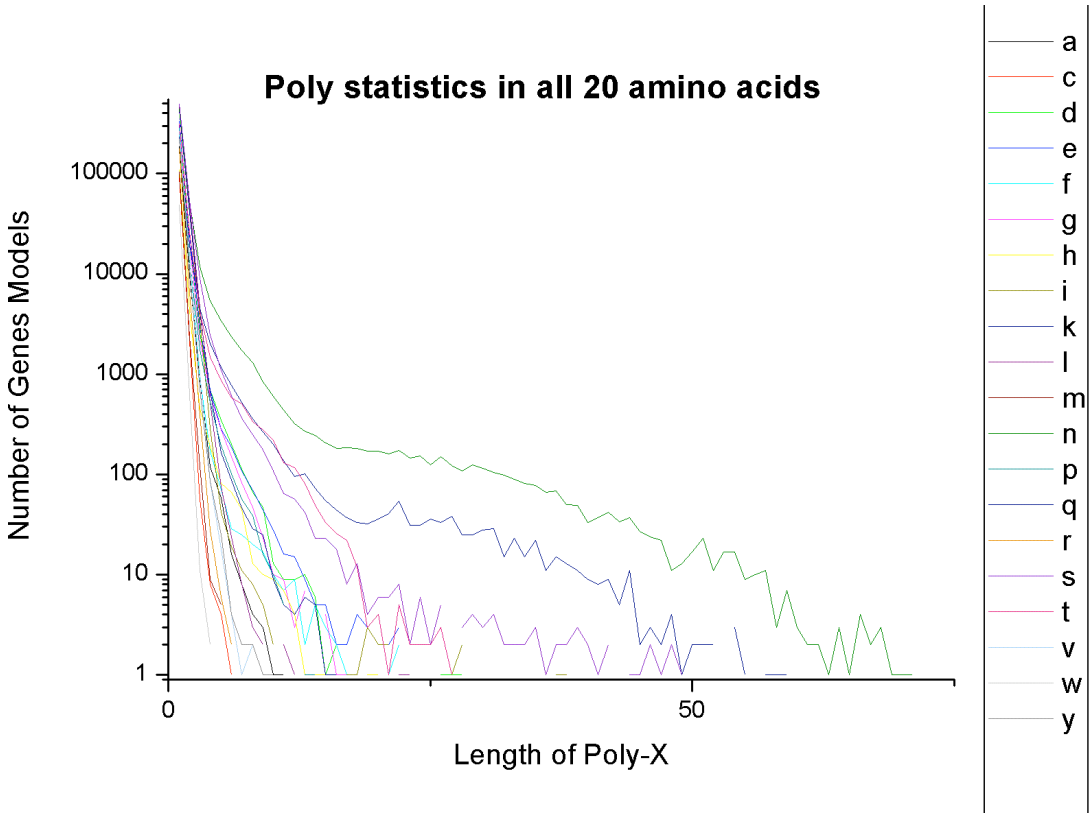
All predicted *Dictyostelium* gene models were translated into peptide sequences and an amino acid composition table for the proteome was generated. Based on that composition, a minimum statistically significant length of repeat for each amino as being non-random is given below at P<0.05, P<0.01, P<0.001. For example, the minimum length of a non-random tract of asparagines (poly-N) is 9 at the P<0.05 level of significance. This is probably a conservative estimate since the existence of extensive poly-N tracts biases the composition of N in the proteome. When compared to yeast, the percentage of A and R are significantly lower, whereas the percentage of N, Q and I are significantly higher in the *Dictyostelium* proteome.

Table SI 5. Distribution of amino acid homopolymers in the proteome.

Amino Acid	Number of tracts	Number of genes
A	35	35
C	6	6
D	274	253
E	211	170
F	105	96
G	187	178
H	154	138
I	18	16
K	86	84
L	14	11
M	4	4
N	5421	2937
P	126	96
Q	2528	1498
R	3	3
S	604	521
T	1302	1079
V	8	8
W	0	0
Y	7	7

All peptide sequences were scanned to identify all homopolymers and their location. They are summarized in 20 mutually exclusive groups although there is considerable overlap in the gene sets that contain different homopolymers. The number of repeats for each amino acid is based on the minimum repeat length at $p<0.01$ in the table above. Altogether, there are 11,095 homopolymer tracts contained in 4,555 genes in *Dictyostelium*. For comparison, there are 345 repeats within 271 genes in the yeast genome that is approximately one-third the size of the *Dictyostelium* genome and has about half the number of genes. Many predicted proteins contain more than one homopolymer tract. There are 2,091 *Dictyostelium* genes that contain polyN and/or polyQ tracts that are ≥20 residues.

Figure SI 7. Distribution of amino acid homopolymers in the predicted proteins of *Dictyostelium*.



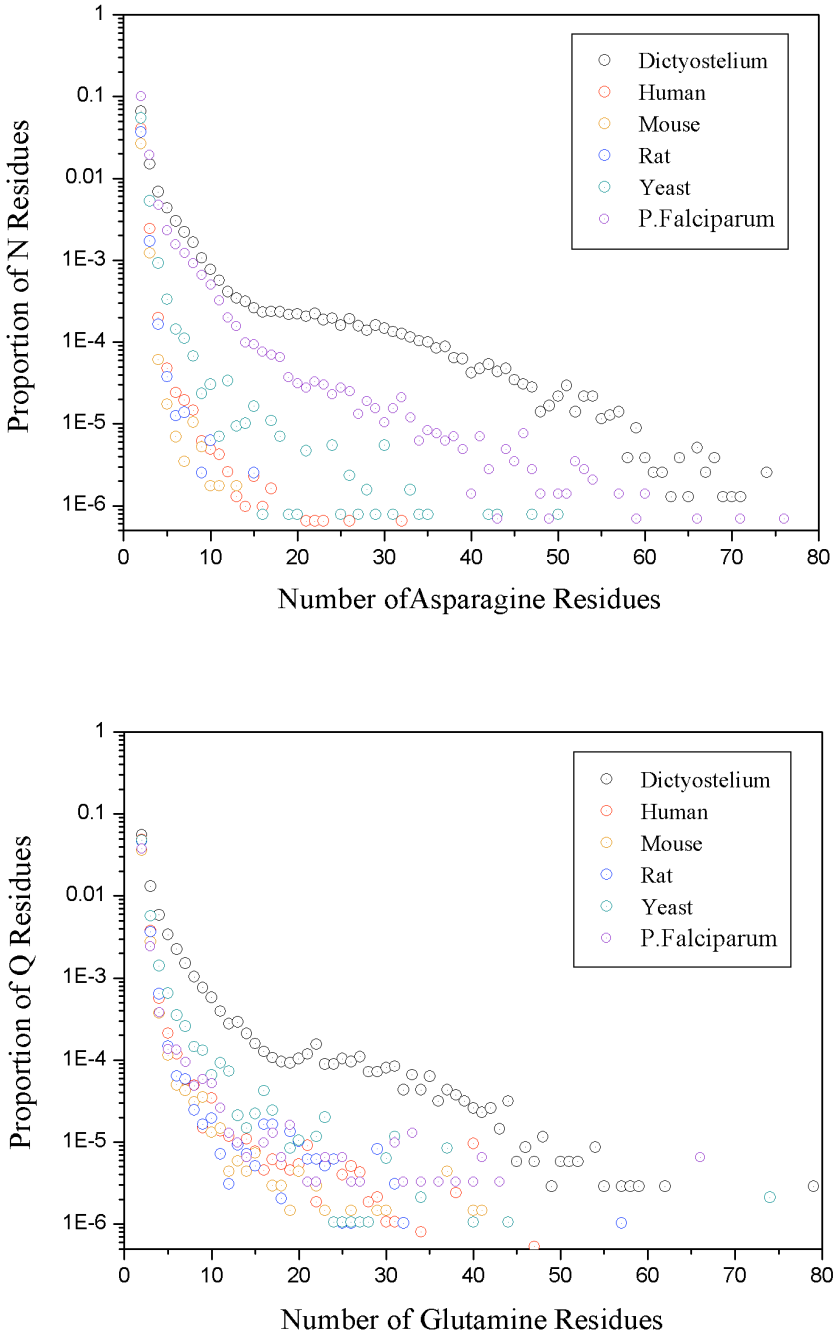
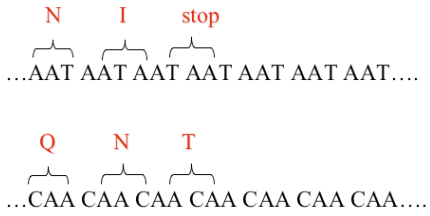


Figure SI 8. Distribution of amino acid homopolymers amongst eukaryotes.
For the predicted proteomes of each of the organisms indicated, the proportion of all asparagine (upper graph) or glutamine (lower graph) residues which lie within homopolymer tracts of the length indicated (horizontal axes) is plotted.

Codon usage in homoploymer tracts

If the homopolymers arose from relatively recent trinucleotide repeat expansions they should be encoded by a single codon rather than a mixture of codons. The serine homopolymers were encoded by each of the six serine codons in roughly equal proportions, after accounting for codon bias, and there were no continuous tracts of a single codon that constituted >30% of any homopolymer tracts >20 residues (data not shown). In contrast, the majority of the Poly-T, Poly-N and Poly-Q tracts were encoded by a single codon.



Of the poly-N tracts >20 residues, 72% were encoded entirely by AAT codons and the rest had uninterrupted runs of AAT that encoded most of the homopolymer interspersed with AAC. Of the Poly-Q tracts larger than 20 residues, 68% were entirely CAA codons and the rest were mostly long runs of CAA interspersed rarely with CAG. These observations suggest that the Poly-N and Poly-Q tracts arose from recent expansions of AAT and CAA triplets. More complicated amino acid repeats are common, such as glutamine-asparagine repeats $(QN)_x$, but they are not encoded by triplet nucleotide repeats. Other than N, AAT tracts can encode isoleucine (I) via the ATA codon. Curiously, the distribution of Poly-I homopolymers in the genome was close to the average of the common amino acids other than S, T, N and Q, suggesting selection against long runs of isoleucine in proteins (Figure SI 7).

Functional Annotation of Poly-N and Poly-Q proteins

To annotate the putative function of selected genes we utilized the Gene Ontology (GO), which is a controlled vocabulary for describing gene function, and analyzed them as described[11,12].The GO annotation for *Dictyostelium* genes were obtained from dictyBase. A recursive traversal of the GO directed acyclic graphs provided the GO identifiers that are overrepresented in the gene list compared to the entire genome. Lists of genes with significantly high representation were identified by comparing the number of genes having a common GO category in the experimental group to the total number of genes having that category in the genome. The data are presented graphically where bar lengths represent the ratio (fold enrichment) between the list frequency (number of genes with a specific GO annotation in list / total number of genes annotated in the list) and the array group frequency (number of genes with that specific GO annotation in the genome / total number of annotated genes in the genome). The x-axis is the scale for that ratio. Connected bars are subgroups of the bars immediately above them, as indicated by the branching pattern. A dotted branch indicates that a group did not show significant enrichment at an intermediate GO level.

P-values were calculated as follows. A value of "0" means that P<0.001.
Given a gene list, we counted the number of genes in the list that map to each GO node in the graph. To estimate the significance of enrichment for a certain GO term, we collected four values:

(1) l is the list node count: the number of genes in the list that are annotated with this term, or any of its subsequent children terms.

(2) m is the list total count: the total number of annotated genes in the list.

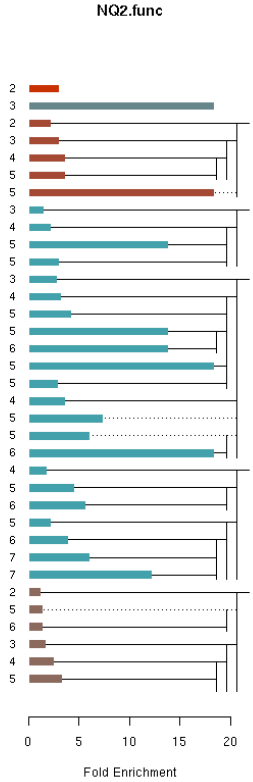
(3) k is the genome node count: similar to the list node count (l), but counts the number of genes in the entire genome instead of the given gene list.

(4) n is the genome total count: the total number of annotated genes in the genome.

The statistical significance is estimated using the hyper-geometric distribution as shown in equation 1, corrected for the problems associated with multiple testing as described[13].

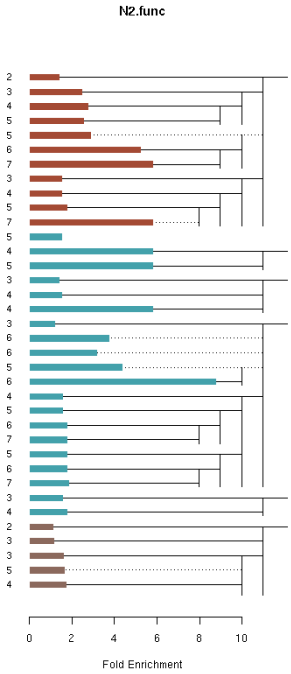
Equation 1. $$P(X \geq l) = 1 - \sum_{i=0}^{l-1} \frac{\binom{k}{i}\binom{n-k}{m-i}}{\binom{n}{m}}$$ Where (k choose i) is defined: $$\binom{k}{i} = \frac{k!}{(k-i)!\,i!}$$

A. The GO "function" annotation distribution for proteins containing Poly-N and Poly-Q.



list	total	p-value	Annotation
13	78	0	transcription regulator activity
3	3	0	two-component response regulator activity
11	89	0.008	enzyme regulator activity
8	48	0.004	GTPase regulator activity
8	40	0.001	small GTPase regulatory/interacting protein activity
6	30	0.005	guanyl-nucleotide exchange factor activity
2	2	0.003	Rab GTPase activator activity
48	574	0.001	transferase activity
41	335	0	transferase activity\, transferring phosphorus-containing groups
6	8	0	phosphotransferase activity\, nitrogenous group as acceptor
40	242	0	phosphotransferase activity\, alcohol group as acceptor
40	254	0	kinase activity
37	208	0	protein kinase activity
10	43	0	protein-tyrosine kinase activity
6	8	0	protein-histidine kinase activity
6	8	0	two-component sensor molecule activity
2	2	0.003	protein threonine/tyrosine kinase activity
14	87	0	protein serine/threonine kinase activity
3	15	0.044	lipid kinase activity
2	5	0.026	phosphoinositide 3-kinase activity
3	9	0.01	phosphatidylinositol phosphate kinase activity
2	2	0.003	phosphatidylinositol-4-phosphate 3-kinase activity
15	147	0.012	hydrolase activity\, acting on ester bonds
4	16	0.009	thiolester hydrolase activity
4	13	0.004	ubiquitin thiolesterase activity
6	48	0.043	nuclease activity
3	14	0.037	endonuclease activity
2	6	0.038	endoribonuclease activity
2	3	0.008	endonuclease activity\, active with either ribo- or deoxyribonucleic acids and pr
83	1255	0.012	binding
21	262	0.043	adenyl nucleotide binding
21	262	0.043	ATP binding
52	550	0	nucleic acid binding
36	255	0	DNA binding
8	44	0.002	transcription factor activity

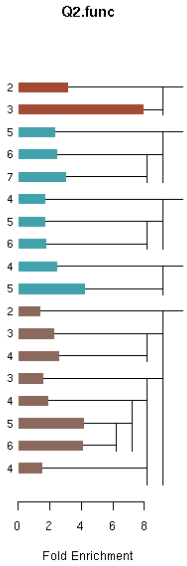
B. The GO "function" annotation distribution for proteins containing only Poly-N.



list	total	p-value	Annotation
51	311	0.003	transporter activity
6	21	0.026	drug transporter activity
6	19	0.016	multidrug transporter activity
5	17	0.036	xenobiotic-transporting ATPase activity
4	12	0.039	ion channel activity
3	5	0.012	voltage-gated ion channel activity
2	3	0.036	voltage-gated chloride channel activity
24	136	0.018	carrier activity
18	102	0.037	primary active transporter activity
18	66	0.009	P-P-bond-hydrolysis-driven transporter activity
2	3	0.036	sodium/potassium-exchanging ATPase activity
43	242	0.001	phosphotransferase activity\, alcohol group as acceptor
2	3	0.036	phosphorus-oxygen lyase activity
2	3	0.036	guanylate cyclase activity
42	254	0.007	kinase activity
37	208	0.003	protein kinase activity
2	3	0.036	diacylglycerol kinase activity
82	583	0.016	hydrolase activity
3	7	0.036	aspartic-type endopeptidase activity
4	11	0.028	mannosidase activity
3	6	0.023	hydrolase activity\, acting on carbon-nitrogen (but not peptide) bonds\, in cyclic amidines
2	2	0.013	deaminase activity
35	194	0.003	hydrolase activity\, acting on acid anhydrides
35	194	0.003	hydrolase activity\, acting on acid anhydrides\, in phosphorus-containing anhydrides
33	159	0	ATPase activity
30	145	0.001	ATPase activity\, coupled
18	87	0.008	hydrolase activity\, acting on acid anhydrides\, catalyzing transmembrane movement of substances
18	87	0.008	ATPase activity\, coupled to transmembrane movement of substances
16	75	0.009	ATP-binding cassette (ABC) transporter activity
15	83	0.045	helicase activity
11	53	0.033	ATP-dependent helicase activity
165	1255	0.007	binding
75	550	0.042	nucleic acid binding
36	195	0.002	metal ion binding
14	73	0.033	zinc ion binding
21	104	0.006	calcium ion binding

(*Figure SI 9 continues overleaf)*

C. The GO "function" annotation distribution for proteins containing only Poly-Q.



list	total	p-value	Annotation
5	35	0.018	motor activity
5	14	0	microtubule motor activity
7	65	0.024	phosphoric ester hydrolase activity
6	54	0.032	phosphoric monoester hydrolase activity
5	36	0.02	phosphoprotein phosphatase activity
15	194	0.025	hydrolase activity\, acting on acid anhydrides
15	194	0.025	hydrolase activity\, acting on acid anhydrides\, in phosphorus
13	159	0.024	ATPase activity
6	53	0.029	ATP-dependent helicase activity
4	21	0.013	ATP-dependent RNA helicase activity
80	1255	0	binding
19	185	0	protein binding
6	51	0.025	cytoskeletal protein binding
40	550	0.001	nucleic acid binding
11	127	0.025	RNA binding
7	37	0.001	mRNA binding
5	27	0.006	pre-mRNA splicing factor activity
18	255	0.033	DNA binding

Figure SI 9. Gene Ontology (GO) Annotation of poly-N and poly-Q homopolymer containing proteins.

Estimates of the over-representation of homopolymer containing proteins in different "function" categories of Gene ontology is shown for proteins containing homopolymer tracts of N or Q (or both) ≥ 20 amino acid residues.

Expression of genes encoding Poly-N and Poly-Q

We examined the expression of genes encoding poly-N and poly-Q tracts since it has been shown that the expression of several $(CAA)_x$ repeat-containing genes is elevated in the first 5 hours of development[14]. Using cDNA microarrays, we could acquire reliable expression data for about 30% of the genes and 59% of them increased in expression at 6-8 hours of development (Figure SI 10). Extrapolating from this data, about 1,200 genes that contain Poly-N and/or Poly-Q tracts appear to be up regulated during development (Table SI 6).

The number of genes and number of up-regulated genes (linear contrast score >0) from Figure SI 10 for each group are listed in Table SI 6. There are 2,091 *Dictyostelium* genes that contain polyN and/or polyQ tracts ≥20 residues. For 646 of them there are targets on the microarray that give significant signals. Since the targets that contain the triplet repeats that encode poly-N or poly-Q are affected by cross-hybridization (see below) we consider only the data from the targets that do not contain repeats. Extrapolating from the numbers of up-regulated genes in Table SI 6, it appears that 59% of these genes (257/438), or 1,227 of them, are up-regulated at 6-8 hours of development.

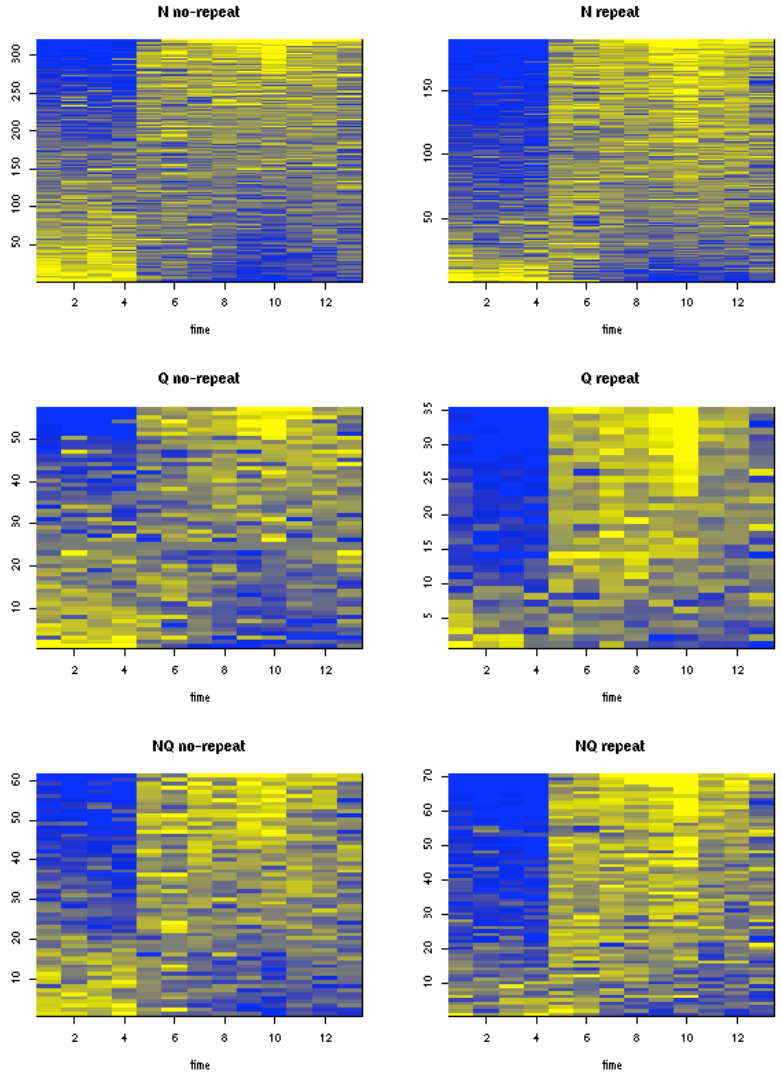


Figure SI 10. Gene expression patterns for poly-N and poly-Q homopolymer proteins. *RNA samples of wild-type cells developing on filters were collected at two-hour intervals and analyzed with a cDNA microarray representing 5,624 genes. The data from the set of poly-N, poly-Q, or poly-N plus poly-Q containing gene targets were plotted to indicate the level of gene expression where the color scale represents the standardized $\log_2$ of the ratio between the test sample and the standard relative to each gene's mean, as described[15]. Blue indicates lower than average level of gene expression for that gene and yellow indicates higher than average level of expression. Each column represents a time-point and each row represents a gene. The time-points labeled 1-13 correspond to 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 and 24 hours of development. The targets are ordered against a linear contrast function as described previously[15]. The three panels on the left are the data obtained from microarray targets that do not contain any of the AAT or CAA repeats that encode the N and Q homo-polymers, while the three panels on the right are the data obtained from targets that include AAT and or CAA repeats. Note that the panels on the right show that nearly all genes represented by the targets are "up regulated" between 6-8 hours of development. This must be due to some level of cross-hybridization of the targets with AAT and CAA repeat-containing cDNAs in the probe because the analogous set of targets without repeats (left-hand panels) show only about one-half of genes are actually up-regulated at this time. See below for test of significance of the difference between Targets containing repeats vs. Targets without repeats.*

Table SI 6. Effect of AAT and CAA repeats on cDNA microarray data.

Homo-polymer Group	Repeats in target?	Genes	Up-regulated Genes (percentage)
N	No	320	183 (57)
N	Yes	189	148 (73)
Q	No	57	32 (56)
Q	Yes	35	29 (82)
NQ	No	61	42 (68)
NQ	Yes	70	55 (79)

Targets on the cDNA microarray are grouped by the homopolymers contained in the genes they represent and by whether they contain the repeats encoding the homopolymers (AAT or CAA) within their sequence that is printed on the array. The sets of genes represented in the "repeat" and "no-repeat" groups only partially overlap.

Tests of statistical significance of the difference between the data obtained from the repeat-containing targets, versus the data from the no-repeat targets.

1. Paired t-test on ratio (up-regulated/total) between two groups.
P=0.06.
Considering that there are only 3 pairs of data, it appears the difference is significant.

2. Chi-square test on N, Q, NQ and combined groups.

N: p=0.004
Q: p=0.12
NQ: p=0.46
combined: p=0.0007.

In summary, both tests show that the ratios of up-regulated genes in repeat groups are higher than in no-repeat groups, which suggests there is some cross-hybridization in repeat groups.

Phylogenetic Tree Construction

The phylogenetic tree in Figure 5 (main text) was constructed in an integrated iterative process described in Olsen and Loomis[16]. The major elements of the process include: (i) construction of a model of orthologous protein sequence divergence; (ii) maximum likelihood estimation of the parameters of this model; (iii) classification of the homology relationships between the protein sequences in the complete or largely complete proteomes of eukaryotic and archaebacterial organisms; (iv) construction of a database of clusters of (likely) orthologous protein sequences, or Evolutionary Clusters of Orthologs (ECOs); (v) serial reconstruction of the phylogenetic tree connecting the organisms represented in the database of ECOs. In this iterative process the number of organisms represented in the database of ECOs and on the organism tree gradually increases, as does the number of ECOs in the database. In comparative performance tests, using protein datasets suitable for re-constructing the ancient divergences in eukaryotic history, the model of orthologous protein sequence divergence out-performed current state of the art phylogenetic models that use Γ distributed rate variation[16]. In addition, the results of bootstrapping strongly confirmed well-accepted parts of the branching topology.

When the predicted proteins encoded by the *Dictyostelium* genome were compared to the proteomes of the 16 organisms, 1,097 of them entered protein clusters with an average of 10.6 members. Thus, each of the clusters that had *Dictyostelium* proteins contained proteins from most of the other proteomes analyzed. Over half the total sequence length of the clustered orthologs could be unequivocally aligned. The tree was rooted on 159 clusters that had representatives from six archaebacterial proteomes (*A. pernix*, *A. fulgidus*, *Halobacterium sp.*, *P. abyssi*, *S. solfataricus*, *T. acidophilum*). The position of *Dictyostelium* was supported by 100 out of 100 bootstraps and a more detailed analysis indicates that the present positioning of *Dictyostelium* would be contradicted, on average, in less than 2 out of 1,000,000 bootstraps (Olsen, unpublished).

The source protein sequence data was downloaded from major archival and genome sequencing institutions. The source files for the 7 archaeal and first 15 eukaryotic proteomes have been described[16]. Additional proteome files were: (i) *Chlamydomonas reinhardtii* (green alga) "fileproteins.finalModelsV2" (dated 2.14.2004) containing a pre-publication version of the entire proteome downloaded from the US Department of Energy Joint Genome Institute (http://genome.jgi-psf.org/chlre1/chlre1.home.html); (ii) a *Giardia lamblia* NCBI Batch Entrez file of 7274 sequences (Entrez search term used "Giardia lamblia") downloaded on 4.6.2004. Out of this set of sequences, 7,160 could be positively identified as *Giardia lamblia* sequences and after culling redundant sequences, 6,804 remained. The *Dictyostelium discoideum* proteome was predicted from Version 2.0 of the genome.

SCOP and Pfam domain distribution

The whole-proteome phylogeny summarizes the relationship of *Dictyostelium* to the major groups of eukaryotes. To explore the potential functional parallels between those groups and *Dictyostelium* we began by defining the protein domains that are shared between them, with an emphasis on those found only in eukaryotic cells. The SUPERFAMILY and Pfam protein domain databases were used to identify protein domains specific to eukaryotes. SCOP superfamilies group protein domains for which there is structural and functional evidence of a single ancestral domain, but they often show little or no sequence similarity, while Pfam family membership is based on sequence and functional similarities[17-20]. SCOP superfamilies group protein domains for which there is structural and functional evidence of a single ancestral

domain, but they often show little or no sequence similarity, while Pfam family membership is based on sequence and functional similarities. The SUPERFAMILY and Pfam databases were first used to identify protein domains specific to bacteria, archaea, and eukaryotes.

(a) Filtering to remove viral and photosynthetic superfamilies

The 1,231 SCOP superfamilies and the 7,316 PFAM families (for simplicity just "superfamilies") were classified as viral if they had one of the following in their name: "viral", "virus", "envelope", "HIV", "hepatitis", "T4", "phage" or "lambda integrase-like".

Superfamilies present in at least one cyanobacterium (Synechocystis sp., Nostoc sp., P. marinus) and at least one photosynthetic eukaryote (A. thaliana and O. sativa, C. reinhardtii, T. pseudonana) but absent from all remaining genomes were classified as photosynthetic. The E-value cutoff here was based on the calculation below but with each genome being considered on its own (rather than using clade-level E-values). Viral and photosynthetic superfamilies were excluded from the rest of the analysis.

(b) E-value scaling

The E-values reported by SUPERFAMILY (and Pfam) are defined for a single query sequence searched against the model library. When assessing the presence or absence of a particular superfamily (or family) in a clade of interest, the query is different: all sequences in that clade against only those models that belong to the superfamily in question. To correct for this discrepancy, we multiplied all E-values by the following factor:

$$S * \frac{L}{300} * \frac{m}{M} * F ,$$

where S is the total number of sequences in the clade, L their average length (amino-acids), m the number of models in a given superfamily, M the total number of models in the library, and F the total number of superfamilies. E-values rescaled in this way give the expected number of superfamilies absent from the clade that are (incorrectly) reported as present due to chance similarities. We used 0.1 as our cut-off for presence.

(c) Kingdom presence/absence

Our dataset consisted of the following 45 genomes:

Eukaryotes (E)	Homo sapiens (Ensembl release 19.34a)
	Fugu rubripes (Ensembl release 19.2)
	Caenorhabditis elegans (Wormbase release WS115)
	Drosophila melanogaster (Flybase release 3.1)
	Neurospora crassa (Broad Inst. release 3)
	Aspergillus nidulans (Broad Inst. release 3.1)
	Schizosaccharomyces pombe
	Saccharomyces cerevisiae
	Oryza sativa (japonica cultivar, TIGR pseudomolecules 1.0)
	Arabidopsis thaliana (TIGR release 5)
	Chlamydomonas reinhardtii (JGI-PSF release 1.0) (*)
	Thalassiosira pseudonana (JGI-PSF) (*)

	Dictyostelium discoideum (Version 2.0)
Archaea	Sulfolobus solfataricus
(A)	Archaeoglobus fulgidus DSM 4304
	Methanothermobacter thermautotrophicus Delta H
	Methanobacterium thermoautotrophicum
	Methanocaldococcus jannaschii
	Methanopyrus kandleri AV19
	Pyrococcus furiosus DSM 3638
	Thermoplasma volcanium
	Nanoarchaeum equitans Kin4-M
Bacteria	Mycobacterium tuberculosis H37Rv
(B)	Bacteroides thetaiotaomicron VPI-5482
	Chlamydophila pneumoniae J138
	Bradyrhizobium japonicum USDA 110
	Bordetella bronchiseptica
	Helicobacter pylori 26695
	Pseudomonas aeruginosa PAO1
	Leptospira interrogans lai 56601
	Mycoplasma genitalium
	Escherichia coli K12
	Onion yellows phytoplasma
	Bacillus anthracis A2012
	Aquifex aeolicus VF5
	Chlorobium tepidum TLS
	Synechocystis PCC 6803
	Deinococcus radiodurans
	Clostridium acetobutylicum
	Geobacter sulfurreducens PCA
	Thermotoga maritima
	Streptomyces coelicolor A3(2)
	Nostoc PCC 7120
	Prochlorococcus marinus marinus CCMP1375
	Lactococcus lactis lactis

(*) These two genomes are available pre-publication from JGI-PSF.

In total there were 13 eukaryotes, 9 archaea and 23 bacteria. The eukaryotic set consisted of 4 animals, 4 fungi and 2 plants, plus one green alga (*C. reinhardtii*, conventinally classified within the plant kingdom), one photosynthetic phytoplankton (*T. pseudonana*, uncertain taxonomy), and *D. discoideum*. The genomes are a taxonomically representative sample chosen from the SUPERFAMILY database (release 1.63). For bacterial and archaeal genomes, preference was given to larger ones, for eukaryotes to model organisms with higher quality protein predictions.

For each of the three kingdoms (Eukarya, Archaea, Bacteria), we classified each superfamily as either: absent from the clade (A); present in a single genome (S); uncertain (U); or widely present in that clade (P).

The criterion for absence was no hit better than the 0.1 cutoff, for E-values scaled by taking all sequences in the kingdom into account (see above). The cutoff is very strict: as explained above, we expect that only 0.1 superfamilies are likely to be classified as not absent based on chance similarity. The converse case -- superfamilies present in the kingdom that are classified as absent because their members have so far avoided detection at the required level of significance -- is difficult to quantify, but the possibility should be kept in mind.

Superfamilies not classified as absent were then subclassified based on the number of genomes in the kingdom in which they were present: single (S) if only in a single genome, widely present (P) if in more than 1/3 of the genomes, and uncertain (U) otherwise. The exact thresholds for wide presence were: >=4 genomes for eukaryotes, 3 for archaea and 7 for bacteria. The kingdom-scaled E-values from the absence calculation were reused in this part, and again the cutoff was 0.1.

The motivation behind the sub-classification step was to account for possible horizontal transfer, in particular from eukaryotes to Archaea and bacteria. The criterion for wide presence is again very strict; only superfamilies that are clearly present in a large number of genomes are counted as widely present.

In the next step, labels from the three kingdoms were combined and defined as present or absent as described in the following table.

Definitions of "Presence" and "Absence" in the kingdoms and combinations of kingdoms.

In the following list, P(A) means "widely present in archaea", S(A) means "present in a single archaeal genome", etc. By doing this, we are assuming that cases where the superfamily is present in only one genome in a particular kingdom are due to horizontal transfer, and that if the superfamily is absent from two kingdoms and uncertain in the third, it is unique to the third kingdom and should be counted towards its total.

ABE	P(A),P(B),P(E)
AE	P(A),A(B),P(E)
	P(A),S(B),P(E)
AB	P(A),P(B),A(E)
	P(A),P(B), S(E)
BE	A(A),P(B), P(E)
	S(A),P(B), P(E)
A	P(A),A(B), A(E)
	P(A),A(B), S(E)
	P(A),S(B), A(E)
	P(A),S(B), S(E)
	U(A),A(B), A(E)

B	A(A),P(B), A(E) A(A),P(B), S(E) S(A),P(B), A(E) S(A),P(B), S(E) A(B),U(B), A(E)
E	A(A),A(B), P(E) A(A),S(B), P(E) S(A),A(B), P(E) S(A),S(B), P(E) A(A),A(B), U(E)
0	A(A), A(B), A(E)
PV	photosynthetic and viral superfamilies
U	all other cases (unclassified)

About 70% of the SCOP superfamily domains could be unambiguously assigned to one or more of the kingdoms, excluding those involved in photosynthesis (Figure SI 11A). Of those, 751 were found in eukaryotes and 261 of these appear to be specific to Eukaryotes. In a similar analysis carried out for Pfam protein families, 1,848 of these families appeared to be specific to eukaryotes (Figure SI 11B). The list of these shared domains is dominated by proteins needed for functions specific to eukaryotic cells such as those involved in cell cycle control, signal transduction, chromatin structure and remodeling, protein glycosylation, the cytoskeleton, vesicular transport, and autophagy.

The eukaryote-specific Superfamily and Pfam protein domains (eSfam's and ePfam's) were then sorted according to their presence or absence within 12 completely sequenced genomes to arrive at their distribution amongst the major groups of organisms, with *Dictyostelium* as the only Amoebozoa (Figure SI 11C & 11D).

Metazoa have retained the highest proportion of all domains, followed by the Plants, Fungi and *Dictyostelium*. Plants, fungi and *Dictyostelium* share a similar proportion of the total number of Pfam domains with metazoa (49, 45 and 40 percent, respectively) and each group shares a distinct set of Pfam domains exclusively with metazoa (132, 66 and 29 domains, respectively). One would expect the proportions for the Amoebozoa to increase as more genomes in this group are completed. The Pfam domains found only in plants and metazoa are predominantly metabolic enzymes; while those specific to plants and fungi are dominated by sugar-handling enzymes (Table SI 7).

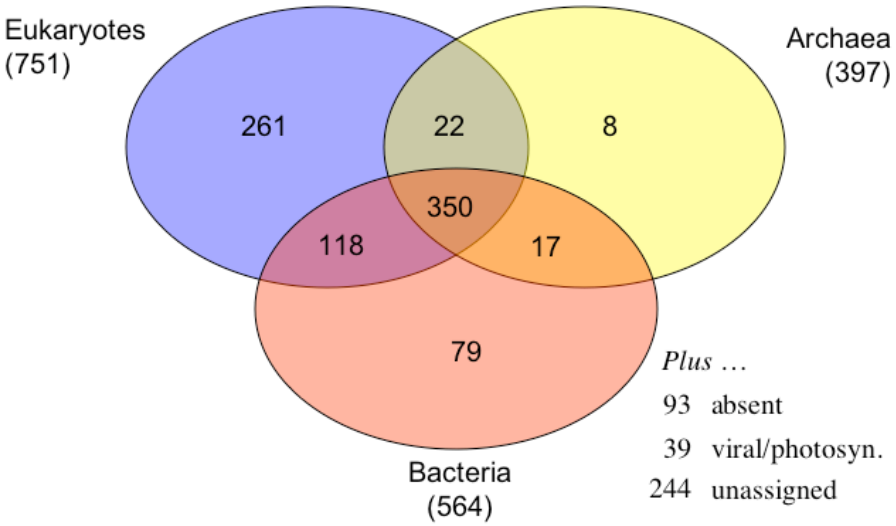


Figure SI 11A. Distribution of the 1,231 SCOP Superfamily domains amongst the kingdoms. *The numbers of domains present in each kingdom or combination of kingdoms are shown.*

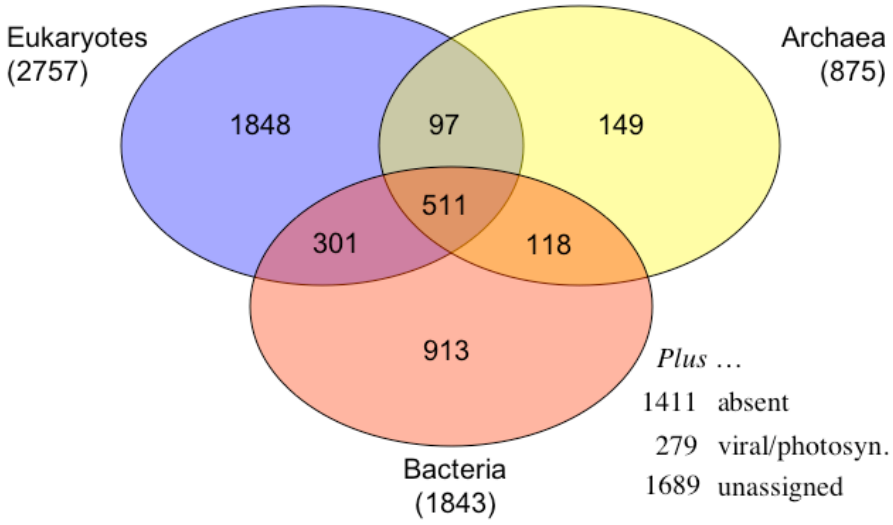


Figure SI 11B. Distribution of the 7,316 Pfam domains amongst the kingdoms. *The numbers of domains present in each kingdom or combination of kingdoms are shown.*

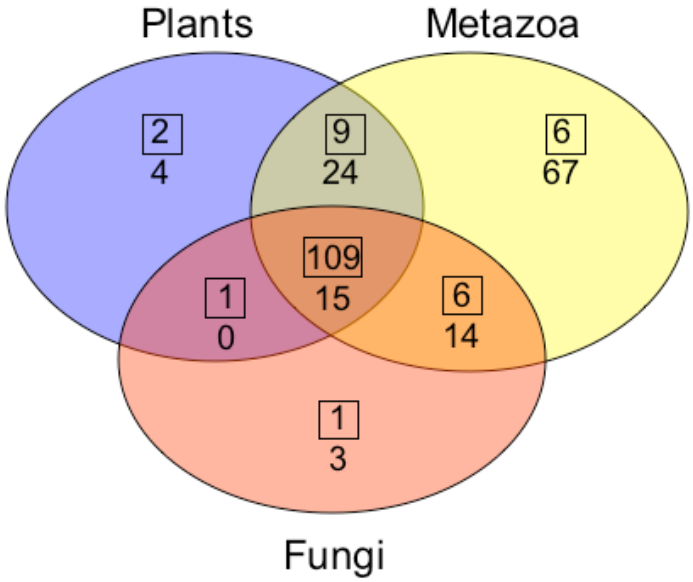


Figure SI 11C. Distribution of eSfam domains amongst the eukaryotes.
The distribution of the 261 eukaryotic-specific SCOP superfamilies (eSfams) amongst the major organismal groups are shown. Boxed numbers represent the number of superfamilies present in Dictyostelium*. The eSfams shared by* Dictyostelium *and other groups of organisms are given in the tables below.*

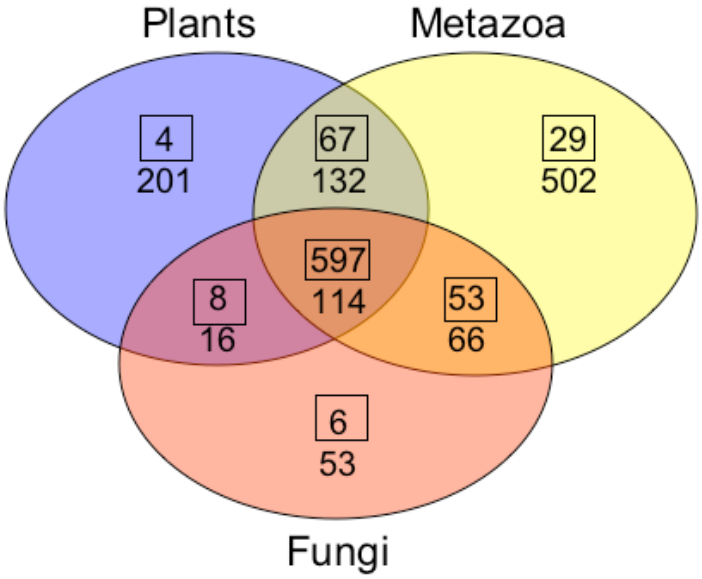


Figure SI 11D. Distribution of ePfam protein families amongst the eukaryotes.
(This is Figure 6 from the main text, duplicated here for comparison.) The distribution of 1,848 eukaryotic-specific Pfam superfamilies (ePfams) amongst the major organismal groups are shown. Boxed numbers represent the number of superfamilies present in Dictyostelium. *The lists of ePfams shared by* Dictyostelium *with other groups of organisms are given in tables below.*

The *Dictyostelium* protein domains shared predominantly or exclusively with plants, fungi or metazoa highlight interesting parallels between *Dictyostelium* and these groups that have functional implications described in the following sections. The protein domains that *Dictyostelium* shares with its sister groups, the Metazoa and Fungi, are interesting because they likely arose soon after plants diverged, but before *Dictyostelium* diverged from the line leading to animals. The major classes of domains include those involved in small and large G-protein signaling (e.g, RGS proteins), cell cycle control and other domains involved in signaling (Tables SI 8 and SI 9). It also appears that glycogen storage and utilization arose as a metabolic strategy soon after the plant/animal divergence since glycogen synthetase seems to have appeared in this evolutionary interval.

Particularly striking are the cases where otherwise ubiquitous domains appear completely absent in one group or another. For instance, *Dictyostelium* appears to have lost the genes that encode collagen, receptor tyrosine kinases, the circadian rhythm control protein Timeless and basic helix-loop-helix transcription factors (Table SI 7). Metazoa, on the other hand, appear to have lost receptor histidine kinases that are common in bacteria, plants and fungi, while *Dicyostelium* has retained and expanded its compliment to 14 members[21]. Fungi have apparently lost copines, which are ubiquitous phospholipids binding proteins, various domains of microfilament system like the villin headpiece and filamin domains and two families of G-protein coupled receptors (see below). The lineage-specific gene losses suggest that a certain degree of functional redundancy must have existed in the early diverging eukaryotes.

Table SI 7. Summary of the distribution of Pfam domains (ePfams) not found in *Dictyostelium*.

Plants Animals and Fungi (114)	
ATP-synt_G	BAG
BAH	CAP
CBM_14	CHORD
Collagen	DNA_pol_delta_4
Exportin-t	GRIM-19
HLH	PAZ
POLO_box	Porin_3
Rad10	Rad21_Rec8
Sec8_exocyst	TIMELESS
TIMELESS_C	zf-RanB
Animals and Fungi (66)	
Arrestin_N	Calcipressin
Chitin_synth_2	Clathrin_lg_ch
CybS	Endosulfine
Fork_head	GPP34
GRIP	HTH_psq
Hemocyanin_M	IATP
Img2	PLA2_B
Peroxin-13_N	RFX_DNA_binding
SIN1	SURF4
Sec2p	Tuberin
Plants and Fungi (16)	
BSP	Caleosin
Chitin_bind_1	Choline_kin_N
DUF1264	DUF455
Glucan_synthase	Glyco_hydro_17
Glyco_hydro_81	Glyco_transf_34
Glyoxal_oxid_N	Isoflavone_redu
Lipase3_N	NAS
Raffinose_syn	Transferase
Plants and Animals (132)	
BRCA2	Cu2_monoox_C
Cu2_monooxygen	Cytochrom_B561
GSH_synthase	Galactosyl_T
Gb3_synth	Glyco_hydro_19
Glyco_hydro_79n	Glyco_transf_29
Glyco_transf_43	Peptidase_M10
Peptidase_M10_N	Prenylcys_lyase
Radial_spoke	Radial_spoke_3
SNAP-25	TNFR_c6
zf-CW	zf-TAZ

A subset of the Pfam domains present are given for each pair, or group, of organisms. The total number of domains is given in parentheses. For the complete listing of ePfam domains see the file "eukaryotic_pfam_detail.txt". See http://www.sanger.ac.uk/Software/Pfam/ for descriptions of each domain.

Table SI 8. SCOP superfamily domains shared amongst eukaryotes and *Dictyostelium*

Number	SCOP Domain Name	SCOP classification
Dictyostelium, animals and fungi		
46966	Spectrin repeat	a.7.1
47912	Wiscott-Aldrich, WASP, C-terminal domain	a.68.1
48065	DBL homology domain (DH-domain)	a.87.1
48366	Ras GEF	a.117.1
57567	Serine proterase inhibitors	g.22.1
81508	Ubiquinone-binding protein QP-C of cytochrome bc1 complex	f.23.13
Dictyostelium, animals and plants		
47031	Second domain of FERM	a.11.2
47050	Thermostable subdomain from villin headpiece	a.14.1
47862	Saposin	a.64.1
54334	Superantigen toxins, C-terminal domain	d.15.6
54403	Cystatin/monellin	d.17.1
55550	SH2 domain	d.93.1
57184	Growth factor receptor domain	g.3.9
57196	EGF/Laminin	g.3.11
81872	BRCA2 helical domain	a.170.1
Dictyostelium, fungi and plants		
54626	Chalcone isomerase*	d.36.1
Dictyostelium and animals		
47216	Proteasome activator reg (alpha)	a.24.8
48670	Transducin, gamma chain	a.137.3
57845	B-box zinc-binding domain	g.43.1
63501	Frizzled cystein-rich domain	a.141.1
81730	beta-catenin-interacting protein ICAT	a.161.1
82927	Cysteine-rich DNA binding domain, (DM domain)	g.62.1
Dictyostelium and fungi		
55154	mRNA triphosphatase CET1	d.63.1
Dictyostelium and plants		
49590	PHL pollen allergen	b.7.3
82653	Probable GTPase Der, C-terminal domain	d.52.5

*Note that Chalcone isomerase is also found in gamma proteobacteria.
For a complete listing of all SCOP SUPERFAMILY domains in other organism categories see the file "eukaryotic_supfam_detail.txt". See http://scop.mrc-lmb.cam.ac.uk/scop/ for a description of the domains.

Table SI 9. Summary of *Dictyostelium* Pfam domains shared with other organisms

Protein Class	Number of distinct Pfam Domains	Predicted function/domain name(s)
Dictyostelium, Animals and Plants (67)		
Cell cycle	4	growth control
Protein biogenesis	4	spliceosome/SRP complex/glycosylation
Cytoskeleton	3	actin binding/bundling, myosin tail
Membrane function	3	Saposin, copine, SCAMP
Cystatin	1	peptidase inhibitors
DOMON (3)	1	cell adhesion
Laminin A	1	basement membrane
Ndr	1	cell differentiation
API5	1	apoptosis inhibitor
Dictyostelium, Animals and Fungi (53)		
Metabolism	8	glycogen syn. ERG2, PEMT, TAPC
Ras/Rho/Rac signaling	7	GAP/GEF/binding
G-protein signaling	3	G-protein regulators
Clatherin function	3	endocytosis
Cell cycle	2	Hus1, Rad1
Signaling	2	Bap31, DAG binding
Cytoskeleton	1	Dynein light intermediate chain
Dictyostelium, Plants and Fungi (8)		
Electron transport	2	ATPase, terminal oxidase
Metabolism	2	sucrase, tRNA phosphoribosyl transferase
Growth	1	RHD3 GTP-binding protein
Dictyostelium and Animals (29)		
GPCR signalling	2	secretin and GABA subfamilies
Signalling	3	PI3K ras BD, PA26, ß-catenin
Cytoskeleton	5	microtubule motor complex, spectrin
Dictyostelium and Fungi (6)		
Transcription	2	fungal Zn cluster, NmrA
Signalling	1	class II cAMP PDE
Transport	1	PDR/CDR ABC transporter domain
DNA repair	1	Mitochondrial genome maintenance
Dictyostelium and Plants (4)		
Regulation	2	WRKY zinc finger

A subset of the Pfam domains present are given for each pair, or group, of organisms that include *Dictyostelium*. The total number of domains is given in parentheses. For the complete listing of ePfam domains see the file "eukaryotic_pfam_detail.txt". See http://www.sanger.ac.uk/Software/Pfam/ for descriptions of each domain.

The analysis of the presence and absence of protein domains in all eukaryotes also identified those that appear to be genuinely *Dictyostelium*-specific (Table SI 10). Since domains are more confidently defined in characterized proteins, it is not surprising that most of these are found in proteins that have been extensively studied. They include a unique family of G-protein coupled receptors that initiate multicellularity by allowing chemotaxis to extracellular cAMP (see below), as well as extracellular matrix proteins and spore coat proteins that form structural elements. In addition, 154 proteins contain 1-18 leucine-rich repeats called FNIP domains that are also commonly found together with b-box zinc finger domains and protein kinase domains. The best studied of the FNIP-domain proteins is Zak1 which is critical in cell fate determination[22].

Table SI 10. Pfam domains so far unique to *Dictyostelium* (of fully sequenced genomes).

Domain Name	Number of Domains	Number of gene models (examples)
Coiled	30	6 (7e, 2c)
Dicty_CAD	4	3 (cadA)
Dicty_CAR	4	4 (carA, carB)
Dicty_CTDC	113	9 (ecmA, ecmB)
Dicty_spore_N	4	4 (cotA, cotB)
FNIP	433	154 (zak1, cigB)
Hisactophilin	1	3 (hatA, hatB)

Additional domain organization information for the FNIP domains can be found at http://www.mrc-lmb.cam.ac.uk/genomes/madanm/dicty/analysis/fnip.shtml.

Analysis of candidates for horizontal gene transfer (HGT) from Prokaryotes

A gene might become stably fixed within a species if it integrates into a totipotent cell, acquires regulatory sequences required for expression and provides a selective advantage to the organism. Amoeba like *Dictyostelium* might have experienced relatively frequent HGT since they are in intimate contact with soil bacteria and since every cell has the potential to give rise to the next generation.

Potential gene transfer events (HGTs) from bacteria were identified by screening for *Dictyostelium* proteins that have a high degree of identity to bacteria-specific Pfam domains and which appear to be absent from all other available eukaryotic genome sequences. However, the use of these criteria alone have led to misinterpretations, in part due to an under-sampling of eukaryotic genomes, so each potential transfer was also examined for phyletic relationships that would be consistent with HGT[23-25].

To search for HGTs from bacteria to *Dictyostelium* a set of bacteria-specific Pfam and Superfamily domains were first identified, using a reference set of fully sequenced bacterial and eukaryotic genomes. The same genomes, searches and statistical methods used to determine the presence and absence of domains, described above, were used for this analysis.

A search for *Dictyostelium* genes related to the bacteria-specific domains identified a number of genes and gene segments with a high degree of identity to bacterial genes and which appear to be absent from all other available eukaryotic genome sequences Using the methods described above for presence and absence of domains, a set of bacteria-specific Pfam and Superfamily domains were defined. We next perfomed Blast searches for sequences with even remote similarity to the proposed HGT sequence (E-value ≤ 0.1). HGT candidates were further checked for absence in other eukaryotes by BLAST searches and examination of the species distribution of the Pfam or Interpro model, ignoring hits to Anopheles, where there is believed to be bacterial contamination of the sequence (http://www.genedb.org/genedb/dicty/index.jsp). Multiple alignments and phylogenetic trees were made with ClustalW. Each potential transfer was examined for phyletic relationships that would be consistent with HGT as described previously[25]. Any potential HGT gene from *Dictyostelium* that clustered within bacterial clades and was not the outgroup of the trees was included in Table 4 (main paper) as a likely gene transfer event. However, we cannot exclude the possibility at this point that these genes are broadly distributed amongst the amoebozoa and have been lost from all other eukaryotes. Although we consider this unlikely, the paucity of complete eukaryotic genome sequences relative to bacterial species makes this a formal possibility. Also, we note that another amoeba, *Hartmannella*, carries the thyA form and not the thyX form of this gene.

Most of the genes have at least one cDNA, which confirmed their intron/exon structures and indicate that they are expressed. The presence of introns was determined by manual inspection using consensus signals. Only the intron in thy1 and several other genes in Table 4 is confirmed by cDNAs. About half of the genes that were not fused to resident genes have one intron and appear fully 'naturalized' in other respects for expression in *Dictyostelium*. The codon usage specifying each domain was not significantly different from other *Dictyostelium* genes, probably due to an adjustment of codon preference after acquisition.

Judging by the top Blast hits amongst the prokaryotes, they appear to have come from different bacterial species and so they likely represent distinct transfer events. Of the proposed HGT examples not cited in the main paper, the isopentenyl transferase domain may provide a function that prevents premature spore germination in that it may make the cytokinin discadenine.

The Cna B domain forms a structural part of a collagen receptor in pathogenic bacteria, such as *Staphylococcus aureus,* that is thought to help present the binding domain away from the cell surface[26]. In *Dictyostelium*, half of the basic dimeric Cna B domain is found embedded within a larger repeated domain of approximately 430 amino acids. The largest Cna B protein, colossin A, is 11,103 amino acids and consists of 91 Cna B segments within 18 of the large repeats, with a novel 500-amino acid C-terminal domain that it shares with colossin B, C, and D (Figure SI 12). Colossin A has a predicted transmembrane domain at its N-terminus, while colossins B, C, and D have predicted signal sequences. The expression of colossin A is restricted to the multicellular stages of development, suggesting that it may form part of the extracellular sheath that envelops the developing organism.

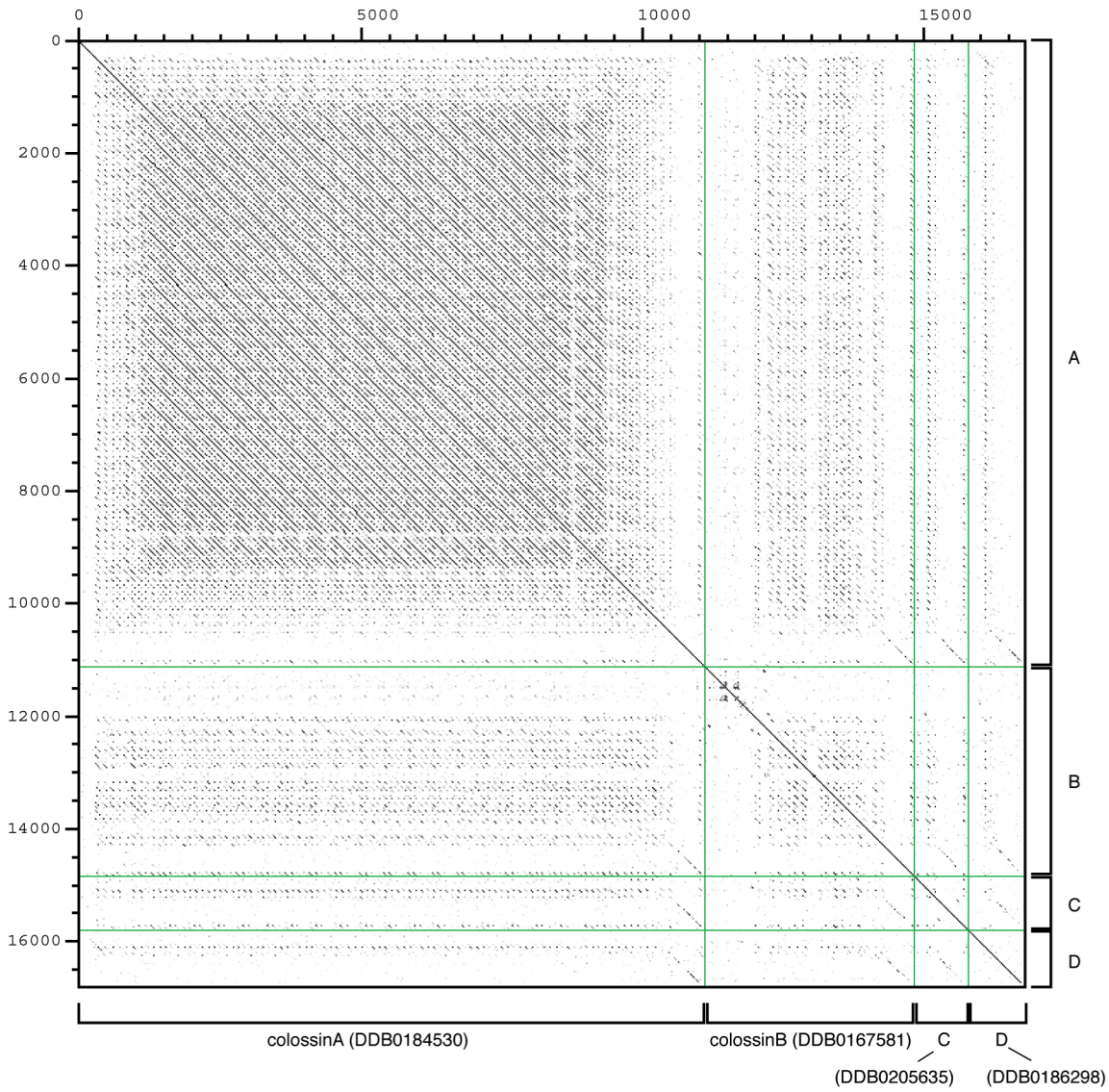


Figure SI 12. The colossin proteins.
An amino acid comparison matrix for the predicted colossin A, B, C and D proteins. The main repeated segments in the center portion of each protein are made up of repeat elements that include the cna B domain, that we propose to have been acquired by HGT, within a larger domain. The C-terminus of each protein contains a shared, non-repeated domain of about roughly 430 amino acids. The dictyBase gene I.D. numbers are given and can be accessed at http://dictybase.org/. Colossin A consists of 91 Cna B segments within 18 of the 430-amino acid repeats.

Each of the potential horizontal gene transfers highlights a unique feature of *Dictyostelium* among eukaryotes, but they may also aid in reconstructing the evolutionary history of Amoebozoa. For example, the amoeba *Hartmannella* has the ThyA form of thymidylate synthase while *Dictyostelium* has the ThyX form[27]. Determining the distribution of the ThyA and ThyX genes, or the other HGTs, should allow a more precise delineation of the phylogenetic relationships amongst the Dictyostelids and perhaps other Amoebozoa, as well.

Polyketide synthases

Many antibiotics and secondary metabolites are produced by polyketide synthases (PKS), modular proteins of around 3,000 amino acids. These enzymes catalyse the repeated condensation of acyl units (normally from malonyl CoA) to give a polyketide, which can be cyclised to form aromatic rings or variously reduced to give hydroxyl or enoyl moieties. We can identify 43 putative PKS genes in *Dictyostelium*, although 10 appear to have arisen by recent duplications on chromosome 5. They account for nearly 2% of the genome's coding potential. In addition, two of the genes have an additional chalcone synthase domain, representing a type of PKS most typical of higher plants and found to be exclusively shared by *Dictyostelium*, fungi and plants in our protein domain comparisons (Table SI 8).

A typical PKS gene consists of the following order of domains from the N-terminus: ketosynthase (condenses malonyl CoA), acyl transferase, dehydratase, methyl-transferase, enoyl-reductase, keto-reductase, acyl carrier protein and thioreductase. In some genes the methyl transferase, enoyl reductase or thio-reductase are missing or are likely inactive. Many of these proteins present a paradox in that they have the capacity to fully reduce the polyketide at each condensation step, which would produce a conventional saturated fatty acid. It is unlikely that so many additional functional genes would have been retained unless they served some other function, especially since the genome contains two highly expressed conventional fatty acid synthase genes. Perhaps the *Dictyostelium* enzymes are able to bypass the reductive loop during some condensation cycles, or they can pass the growing polyketide to another PKS, which lacks a full reductive loop. We cannot recognize dimerization domains in the *Dictyostelium* proteins, analogous to those of the bacterial proteins, so the latter possibility seems unlikely. Since there is no known mechanism for bypassing the reductive loop, a novel mechanism appears to be at work.

Analysis of cellulose metabolism genes and predicted proteins

Dictyostelium cells, like those of metazoa, become strongly adhesive to each other during development so that they can be organized into tissues. To assist the organization of cells into tissues and the tissues into a fruiting body, *Dictyostelium* has evolved both an extra-cellular matrix-the slime sheath-and a skeletal element-the stalk tube. Cellulose is a component of the sheath that surrounds the cell aggregates that form during development and cellulose is deposited in the stalk, stalk cell walls and spore coats[28-30]. So, unlike metazoa that use collagen and keratin to construct analogous structures, *Dictyostelium* bases its structural elements on cellulose, and the proteome reveals an impressive array of proteins by which cellulose can be made and handled. Many other proteins are organized into these structural elements, including ones that bind cellulose directly, and conceivably the CnaB-repeat proteins (the colossins) recruited from bacteria.

The *Dictyostelium* genome carries at least 40 genes whose products are likely to be involved in cellulose synthesis or degradation. Cellulose is synthesized from UDP-glucose and degraded in a multistep process that involved mercerization of the crystals followed by hydrolysis of the polymers by three types of hydrolytic enzymes. Cellulases cut internal beta-1,4-glucosidic bonds, then exocellobiohydrolases attack the non-reducing ends of the cellulose polymer chains releasing the dissaccharide cellobiose which is hydrolyzed by beta-1,4-glucosidases releasing glucose. *Dictyostelium* has all of the enzymes needed for these reactions as well as other cellulose binding proteins.

The first eukaryotic cellulose synthase was discovered in a mutant of *Dictyostelium* (*dcsA*, DDB0190533) that failed to form spores or stalks[31]. This enzyme has now been recognized in many plants as well as the fungus *Neurosporra crassa* and the ascidian *Ciona intestinalis*[32]. The fungal and urochordate enzymes are more closely related to the *Dictyostelium* homolog than to bacterial cellulose synthases indicating that the common ancestor of fungi and animals carried a gene for cellulose synthase that was subsequently lost in most animals. Classical plant cellulose synthases are more distantly related and cluster with the cellulose synthase from cyanobacteria and of many other bacteria. There is no indication of a recent gene transfer to explain the presence of cellulose synthase in *Dictyostelium* or *Ciona.*

Glycosyl hydrolases and expansins

Dictyostelium carries 7 members of the endoglucanase family 9 as well as two pseudogenes of this family (*celA-G*). One member of this family, CelA, has been biochemically shown to be involved in cellulose degradation during germination of spores[33]. The other 6 putative cellulases are clear paralogs to *celA* and are about equally distant from plant, bacterial, fungal and animal cellulases. They are found in two clusters on chromosome 4 suggesting that they arose fairly recently by tandem duplication.

The family 5 of glycosyl hydrolase are found mostly in bacteria that degrade wood and are distantly related to the cellulases of family 9 endoglycosyl hydrolases. There are four *Dictyostelium* genes that cluster with this family.

A related family of genes encode the expansins that fascilitate disruption of non-covalent bonds in plant cell walls. Five members of this family were recently described in *Dictyostelium*[34]. Inspection of the whole genome showed that there is one more member of this family. While clearly related to the plant expansin genes, the *Dictyostelium* proteins cluster with a cyanobacterial expansin protein and a domain from a bacterial glycosyl hydrolase 9 showing the ancient evolutionary origin of these genes.

Lichenase, Xylenase and cellobiohydrolase

Lichenin is a polymer of mixed 1,3-1,4-beta-D-glucans that is degraded by specialized enzymes called lichenases or licheninases. There are 4 genes in the *Dictyostelium* genome that encode proteins distantly related to this family.

Xylan polymers, often found associated with cellulose in higher plants, are degraded by xylenases. *Dictyostelium* has a single putative xylenase with a classical secretion signal. It shows greater than 40 % identity with bacterial xylenase sequences but appears unrelated to xylanesases of eukaryotes. This gene is another candidate for lateral transfer from a bacterial genome following the radiation of the crown organisms.

The non-reducing end of a cellulose polymer is attacked by cellobiohydrolases of the family 7 of glycosyl hydrolases that release the disaccharide cellobiose. A single putative *Dictyostelium* gene shares more than 60% identity with proteins of this family (cbhA, DDB0189810). It was previously thought that family 7 was restricted to fungi but it appears to be an older family. A partial sequence from a mussel indicates that cellobiohydrolases may be widespread.

Cellulose binding proteins

At least 21 cellulose binding proteins without obvious enzymatic domains can be identified in the *Dictyostelium* genome. The products of several of these genes (*celB*, *shnC*, *shnD*, St15, *staB* and *pspB*) have been directly shown to bind cellulose or hemicellulose[30, 35, 36]. Most

Dictyostelium cellulose binding proteins are small and contain little more than a signal sequence for secretion and one to three binding domains that are similar to those in family 9 glycosyl hydrolases of plants. Many of the genes that cluster together on the basis of sequence homology are found adjacent in the genome indicating that they arose by tandem duplication.

Proteins of the actin cytoskeleton and upstream regulators

Analysis of the presence or absence of distinct actin binding proteins in plants, fungi, and metazoa revealed that the *Dictyostelium* repertoire of cytoskeletal proteins is most similar to metazoa followed by fungi (Table SI 11). However, a number of actin-binding proteins, like comitin and ponticulin, are so far unique to *Dictyostelium*. Surprisingly, although the actin cytoskeleton has been studied for over twenty-five years, 71 actin-binding proteins apparently escaped classical methods of discovery. For example, actobindins had not been previously recognized in *Dictyostelium.* Some domains like the actin depolymerisation factor (ADF) domain and the calponin homology (CH) domain appear to have expanded followed by diversification and domain shuffling (Table SI 12). There are thirteen genes that encode ADF domains, including members of the cofilin, coactosin, twinfilin and glia maturation factor subfamilies (Figure SI 13). The CH domain family can be subdivided into several subfamilies according to the type and arrangement of the CH domains. Curiously, a substantial fraction of CH domain proteins have domain combinations unique to *Dictyostelium*. For the rest of the CH domain proteins there are hardly any counterparts in plants or fungi, but there appear to be orthologs for at least one member of most subfamilies in *C. elegans*, *D. melanogaster* or vertebrates. We also identified 11 genes encoding actin related proteins (ARPs) of which three might be founding members of a new class (Figure SI 14). There are apparent orthologues of all ARP classes present in mammals, but no ARP7 or ARP9 proteins that are found exclusively in *S. cerevisiae*. The genome also encodes an unusual member of the actin gene family, filactin, which is unique amongst the sequenced genomes in that it has two N-terminal Ig repeats followed by a conventional actin domain.

This genome-wide survey of the microfilament system strongly supports the concept of functional redundancy, which proposes that the cytoskeletal network is composed of overlapping activities that can functionally compensate one another[37]. Apart from notable exceptions such as the Arp2/3 complex, many proteins of the actin cytoskeleton are encoded by two or more genes. Alternatively, less related proteins appear to exhibit similar activities (Figure 7, main paper; Tables SI 11 and SI 12). A rather extreme example for redundancy is the actin gene family itself, where 17 encode identical proteins with more than 94% identity to human β-actin and 13 encode proteins that are 58-94% identical. This large number of very similar actin proteins probably reflects the need for functionally identical actin proteins to be produced at different times during development and in different tissues, because nearly identical actin genes often show radically different patterns of expression[38-41].

Table SI 11. ***Dictyostelium*** **actin-interacting proteins and their occurrence in other phyla** ***(table continues overleaf)***

Protein Class	Number (new)	Actin-binding module	Closest relatives in other organisms	*Occurence*
Monomeric actin binding				
Profilin	3 (1)	Profilin fold	Profilin	E
Actobindin-like	3 (3)	WH2	Actobindin, Thymosin β4	P, M
CAP	1	WH2	CAP/ASP56	E
WH2-containing[1]	5 (5)	WH2	Unique	U
Twinfilin-like	1 (1)	ADF	Twinfilin	F, M
Capping and/or severing				
Cap32/34 (Aginactin)[2]	2	Cap fold	CapZ	E
Cofilin	6 (4)	ADF	Cofilin/ADF	E
Severin	1	GEL	CapG, Fragmin	M
GRP125	1	GEL	Gelsolin	U
Gelsolin-related[3]	2 (2)	GEL	Gelsolin	M
Capping and nucleation				
Arp2/3 complex[2]	7	Actin fold	Arp2/3 complex	E
Scar	1	WH2	WAVE	M
WASP	1 (1)	WH2	WASP	F, M
WASP-related[4]	2 (2)	WH2	WASP/WAVE (Unique)	U
VASP	1	EVH2	Ena/VASP	M
Formin[5]	10 (1)	FH2	Formins	E
Cross-linking				
ABP34	1		Unique	U
eEF1A (ABP50)	2	eEF1A fold	eEF1A	E
eEF1B	3 (2)		eEF1B	E
Dynacortin	1		Unique	U
Fimbrin	1	CH[6]	Fimbrin/Plastin	E
Fimbrin type ABD-containing[6]	5 (5)	CH	Unique	U
Filamin (Gelation factor, ABP120)	1	CH	Filamin	M
α-actinin	1	CH	α-actinin	F, M
Cortexillin	2	CH	Unique	U
α-actinin type ABD-containing[6]	3 (3)	CH	Unique	U
Protovillin (Cap100)	1	GEL, VHP	Villin	P, M
Villin-related[7]	1 (1)	GEL, VHP	Villin (Unique)	U
Flightless/Villin-related[8]	1 (1)	GEL, VHP	Flightles, Villin (Unique)	U
Villidin[9]	1	GEL, VHP	ABPH, Coronin, Villin (Unique)	U
Kelch-related[10]	1(1)	KELCH	Kelch, Mayven	M
Lateral actin binding				
Smoothelin-related	1 (1)	CH	Smoothelin	M
GAS2-related	1 (1)	CH	GAS2	M
CH-containing[6]	19 (18)	CH	Mostly unique	U
VHP-containing	3 (3)	VHP	Unique	U
Coronin	1		Coronin	F, M
Coronin-like	1 (1)		Coronin7/POD	M
Aip	1		Aip	E

Coactosin	1	ADF	Coactosin, Drebrin	F, M
Coactosin-related[11]	3 (3)	ADF	Coactosin, Drebrin (Unique)	U
Abp1	1 (1)	ADF	Abp1/DrebrinF	F, M
Glia maturation factor-related	1 (1)	ADF	GMF	M
LIM domain-containing[12]	3		LIM proteins (Unique)	U
Membrane-associated				
Interaptin	1	CH	Syne/Anc-1	M
Ponticulin	2		Unique	U
Ponticulin-related	2 (2)		Unique	U
Comitin	1		Unique	U
Comitin-related	1 (1)		Unique	U
Hisactophilin	3 (1)	Trefoil fold	Fascin (Unique)	U
Talin A (Filopodin)	1	I/LWEQ,	Talin	M
Talin B	1	I/LWEQ, VHP	Talin (Unique)	U
SLA-2-like	1 (1)	I/LWEQ	SLA2/HIP1	F, M
Annexin	2 (1)		Annexin	E
Vinculin/α-catenin-related	2 (2)		Vinculin, α-catenin	M
Motors				
Conventional myosin	1	MYO	Conventional myosin	F, M
Unconventional myosins[13]	12 (1)	MYO	Unconventional myosins	E

The proteins have been classified primarily according to their most prominent activity and each protein is displayed only once in the table. However, it should be noted that many proteins display more than one activity. Those proteins for which no biochemical data were available have been grouped according to their structural relationship to characterised proteins in *Dictyostelium* or to orthologues in other organisms, therefore their position in the table should be taken as putative.
Abbreviations: **ADF**, actin depolymerisation factor/cofilin-like domain; **CH**, calponin homology domain; **EVH2**, Ena/VASP homology domain 2; **FH2**, formin homology 2 domain; **GEL**, gelsolin repeat domain; **I/LWEQ**, actin-binding domain of talin and related proteins; **KELCH**, Kelch repeat domain; **MYO**, myosin motor domain; **VHP**, villin head piece; **WH2**, Wiskott Aldrich syndrome homology region 2; **U**, unique (protein has no relatives in plants, fungi or metazoa, or differs from relatives due to extensions or an unusual domain composition; two cases of true homologues in protists, *P. pallidum* fragmin and *E. histolytica* ABPH are mentioned); **E**, eukaryotes; **P**, plants; **F**, fungi; **M**, metazoa (for E, F, P and M the protein might be missing from any particular species).
[1]Four of the WH2-containing proteins bear some relationship with the N-terminus of verprolins and WASP-interacting protein.
[2] Each subunit is encoded by a single gene.
[3] N- and/or C-terminal extensions; one of the proteins contains two Rho GTPase-binding domains which are not present in other gelsolin-like proteins.
[4] Rho GTPase-binding domain, Proline-rich and WH2 domains in common with WASP, but lack WH1 domain.
[5] Except for formin C, which lacks an FH1 region, all other formins have the common FH1-FH2 structure; formins A, E and J have additional domains.
[6] Detailed description of proteins with CH domains in Table SI 12.
[7] N-terminal coiled-coil extension.
[8] Fusion of a truncated flightless and a villin-like protein.
[9] Villidin appears to be a fusion of a coronin and a villin-like protein.
[10] The *Dictyostelium* genome encodes at least 30 proteins with kelch repeats, but actin-binding properties are well documented only for the family of Kelch-related proteins. The putative *Dictyostelium* homolog has not been characterized yet.
[11] N-terminal extension or duplications of ADF domains.
[12] The *Dictyostelium* genome encodes at least 24 proteins with LIM domains, but actin-binding properties are well documented only for three of them, LimC, LimD and LimE.
[13] There are clear representatives of myosin classes I and VII, but other members might constitute unique classes.

Although closer to metazoa and fungi, there are notable absences in the *Dictyostelium* actin cytoskeleton. These include monomeric actin-binding proteins like DNAse I and Vitamin D-binding protein, membrane anchors like MARCKS, tensin, synapsin, band4.1 and members of the ezrin/radixin/moesin family, as well as crosslinkers like fascin, dematin, espin and scruin, all present in metazoa. Septin and anillin, actin crosslinkers found in metazoa and fungi, are also absent in *Dictyostelium*. Among the activators of the Arp2/3 complex cortactin and Pan1p, described in metazoa and fungi, respectively, are not found in *Dictyostelium*, but the multidomain scaffold protein CARMIL, also found in metazoa, is present. Not surprisingly, a large number of proteins that organize the cytoskeleton of the muscle cells are not found in *Dictyostelium*. These include troponin, tropomodulin, caldesmon, adducin, nebulin and titin. We cannot exclude that a distant relative of tropomyosin, a protein that binds along actin filaments in fungi and metazoa, will be found in the future. The coiled-coil composition of this protein makes similarity searches unreliable.

Remodeling of the actin cytoskeleton in response to chemoattractants and during phagocytosis is regulated by signaling pathways that to a large extent make use of small GTPases of the Rho family (Table SI 13). Like in the case of the cytoskeletal proteins, the repertoire of Rho signaling components is more similar to metazoa and fungi than plants. Of the 15 Rho family GTPases in *Dictyostelium*, some are clear Rac orthologues and one belongs to the RhoBTB subfamily. However, the Cdc42 and Rho subfamilies characteristic of metazoa and fungi are absent. The activities of these GTPases are regulated by two members of the RhoGDI family, by components of ELMO1/DOCK180 complexes and by a surprisingly large number of proteins carrying RhoGEF and RhoGAP domains (>40 of each), most of which show domain compositions not found in other organisms. Among the (in many cases putative) effectors found in *Dictyostelium* are the CRIB domain proteins (WASP and related proteins, 8 PAK kinases and a novel gelsolin-related protein), components of the Scar/WAVE complex, formins, IQGAPs, lipid kinases, phospholipases, NADPH oxidase and CIP4. Remarkably, *Dictyostelium* appears to be the only lower eukaryote that possesses class I PI 3-kinases, which are at the crossroad of several critical signalling pathways[42]. By contrast, Rho-specific effectors like Rhotekin, Rhophilin, PKN and ROCK are apparently missing, consistent with the absence of members of the Rho subfamily in *Dictyostelium*. The diverse array of these regulators and the discovery of many additional actin-binding proteins suggest that there are many aspects of cytoskeletal regulation that have yet to be explored.

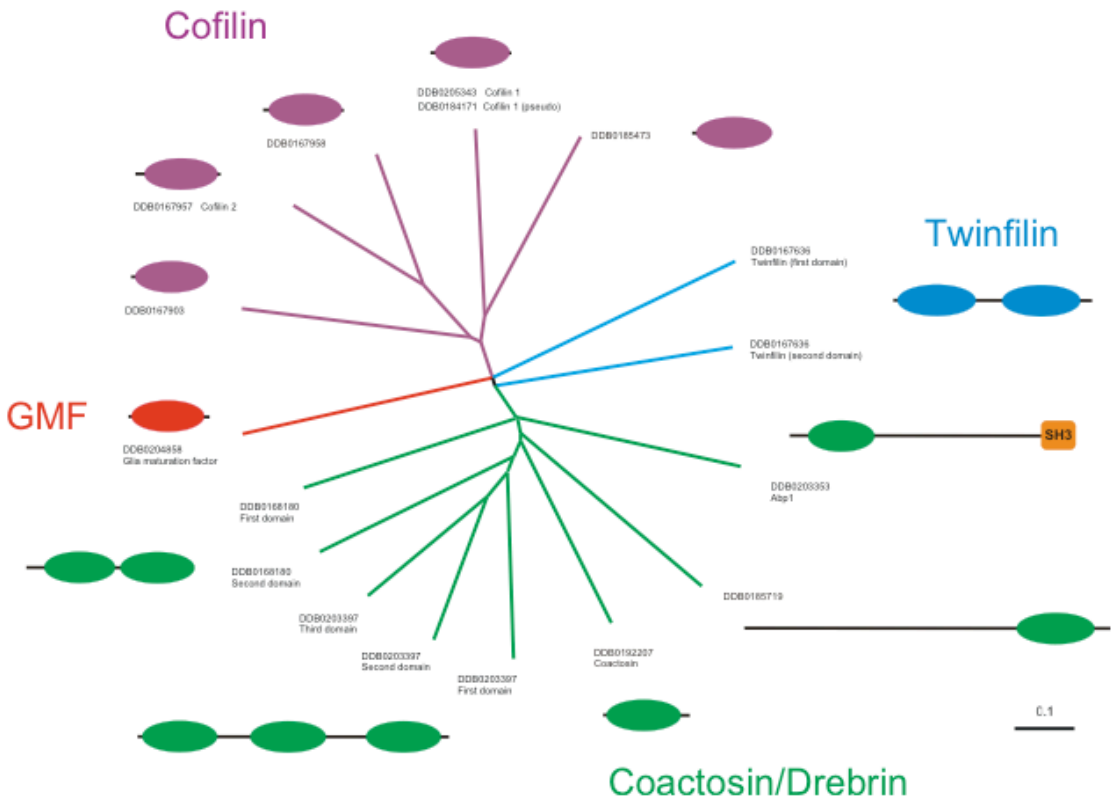


Figure SI 13 ADF domain containing proteins in *Dictyostelium*.
A CLUSTALX alignment of the actin depolymerisation factor (ADF) domains in the Dictyostelium *genome was used to create an unrooted dendrogram with the TreeView program. Major subfamilies and the ADF domains have been colour coded and the domain composition of each protein is depicted. The DDB gene identifiers can be used to locate the annotation for that gene model at dictyBase (http://dictbase.org/) or at geneDB (http://www.genedb.org/genedb/dicty/index.jsp)*

.

Table SI 12. The family of CH domain proteins in *Dictyostelium*. (*Continues overleaf*)

Gene identifier (identity)	Domain structure	Closest relatives	Occurrence					Notes
			P	F	Ce	Dm	V	
Fimbrin-like								
DDB0204382 (fimbrin)	2xEFh-CHf1-CHf2-CHf3-CHf4	fimbrin/plastin	Y	Y	Y	Y	Y	
DDB0169505	PH-coil-CHf1-CHf2-CHf3-CHf4	fimbrin/plastin	(Y)	(Y)	(Y)	(Y)	(Y)	1
DDB0205524	coil-CHf1-CHf2-CHf3-CHf4	fimbrin/plastin	(Y)	(Y)	(Y)	(Y)	(Y)	1
DDB0201990	coil-CHf1-CHf2-CHf3-CHf4	fimbrin/plastin	(Y)	(Y)	(Y)	(Y)	(Y)	1
DDB0184203	CHf1-CHf2-coil		N	N	N	N	N	
DDB0202463	CHf1-CHf2-RasGAP-RasGAP_C	IQGAP	N	(Y)	(Y)	N	(Y)	2
CH1-CH2								
DDB0187455 (interaptin)	CH1-CH2-coil-TM	Syne/ANC-1	N	N	(Y)	(Y)	(Y)	3
DDB0190932 (filamin)	CH1-CH2-6xIG_FLMN	filamin	N	N	Y	Y	Y	
DDB0190048 (α-actinin)	CH1-CH2-4xSPEC-2xEFh	α-actinin	N	Y	Y	Y	Y	4
DDB0188442 (cortexillin 1)	CH1-CH2-coil		N	N	N	N	N	
DDB0168301 (cortexillin 2)	CH1-CH2-coil		N	N	N	N	N	
DDB0218075	CH1-CH2		N	N	N	N	N	
DDB0218148	CH1-CH2-IQ-RhoGEF-PH		N	N	N	N	N	5
DDB0219268	RA-PH-CH1-CH2-RA		N	N	N	N	N	6
CH1								
DDB0185859	CH1	NAV/steerin/ helicase/UNC-53	N	N	Y	N	Y	
DDB0204946	CH1-IQ-RhoGEF-PX		N	N	N	N	N	5
DDB0218183	CH1-2xIQ-RhoGEF-PH-PH-ArfGAP-PH		N	N	N	N	N	5
CH2								
DDB0206046	CH2-coil		N	N	N	N	N	
DDB0185522	CH2-coil		N	N	N	N	N	
DDB0217023	coil-CH2	smoothelin	N	N	Y	Y	Y	
CH3								
DDB0190886	CH3-13xANK		N	N	N	N	N	
DDB0189321	CH3-6xLIM		N	N	N	N	N	7
DDB0190249	CH3-CH3-CH1-RhoGEF-PH		N	N	N	N	N	5
DDB0189592	CH3-IQ-RhoGEF-PH-VHP		N	N	N	N	N	5
DDB0192204	CH3-IQ-RhoGEF-PH-VHP		N	N	N	N	N	5
DDB0169060	CH3-IQ-RhoGEF-PH		N	N	N	N	N	5
DDB0204350 (RacGEF 1)	CH3-IQ-RhoGEF-PH		N	N	N	N	N	5
DDB0204611 (RasGEF P)	CH3-RasGEFN-RasGEF		N	N	N	N	N	
DDB0190478	CH3-C1-PBD-S_TKc		N	N	N	N	N	
DDB0188227	CH3-GAS2	GAS2	N	N	N	Y	Y	
DDB0184511	CH3-C1		N	N	N	N	N	
DDB0190592	CH3		N	N	N	N	N	
DDB0189688	CH3		N	N	N	N	N	
DDB0204997	CH3-TM		N	N	N	N	N	8
CHe								
DDB0218515 (EB1)	CHe-coil	EB1	Y	Y	Y	Y	Y	

CH domain containing proteins have been classified by the sub-class of CH domains they contain and by the other domains that are present. Additional domains are abbreviated as follows: ANK, ankyrin

repeat; ArfGAP, Arf GTPase activating protein; C1, Protein kinase C conserved region 1; CR, calponin repeats; EFh, calcium-binding EF hands; GAS2, growth-arrest-specific protein 2 domain; IG_FLMN, filamin-type immunoglobulin domain; IQ, calmodulin-binding motif; LIM, zinc-binding domain; PBD, p21(Rho)-binding domain; PH, pleckstrin homology domain; PX, Phox homologous domain; RA, Ras association domain; RasGAP, Ras GTPase activating protein; RasGAP_C, RasGAP C-terminus; RasGEF, Ras GTPase guanine nucleotide exchange factor; RasGEFN, rasGEF N-terminal motif; RhoGEF, Rho GTPase guanine nucleotide exchange factor; SPEC, spectrin repeat; S_TKc, Ser/Thr protein kinase, catalytic domain; TM, transmembrane region; VHP, villin head piece. Coiled-coil regions have been included only where extended or functionally relevant. Occurence refers to: P, plants; F, fungi; Ce, *C. elegans*; Dm, *D. melanogaster*; V, Vertebrates.

1. Fimbrin-related proteins are devoid of EF hands and possess long N-terminal extensions not present in any other known fimbrin.
2. Resembles IQGAP, although it lacks IQ repeats and has two CH domains of the fimbrin type instead of one of the CH3 type.
3. Interaptin shares with Syne/Anc the presence of an actin-binding domain and a nuclear envelope-targeting transmembrane region separated by a long central region which in interaptin is predominantly of coiled coil structure.
4. *Dictyostelium* lacks further members with spectrin repeats, like spectrin, dystrophin, plakins; the Syne equivalent (interaptin) lacks clear spectrin repeats.
5. The combination CH-RhoGEF as appears in *Dictyostelium* is unique. In other organisms a CH3-RhoGEF combination is present in Vav, α-PIX and Cdc24; characteristic of *Dictyostelium* CH-RhoGEFs are: a) CH domains not only of type 3, and sometimes more than one, b) IQ domains (usually one), in almost all cases, c) additional domains, like VHP, PX, ArfGAP.
6. The RA-PH combination at the N-terminus is present in metazoa in a number of proteins, but the combination of RA with CH appears unique to *Dictyostelium*.
7. The combination CH + LIM exists in metazoa in a number of proteins, but these have only one LIM domain.
8. The closest relatives are leucine-rich repeat neuronal proteins of metazoans, with the domain composition LRR-CH3-TM, but in these the TM region is very close to the CH domain.

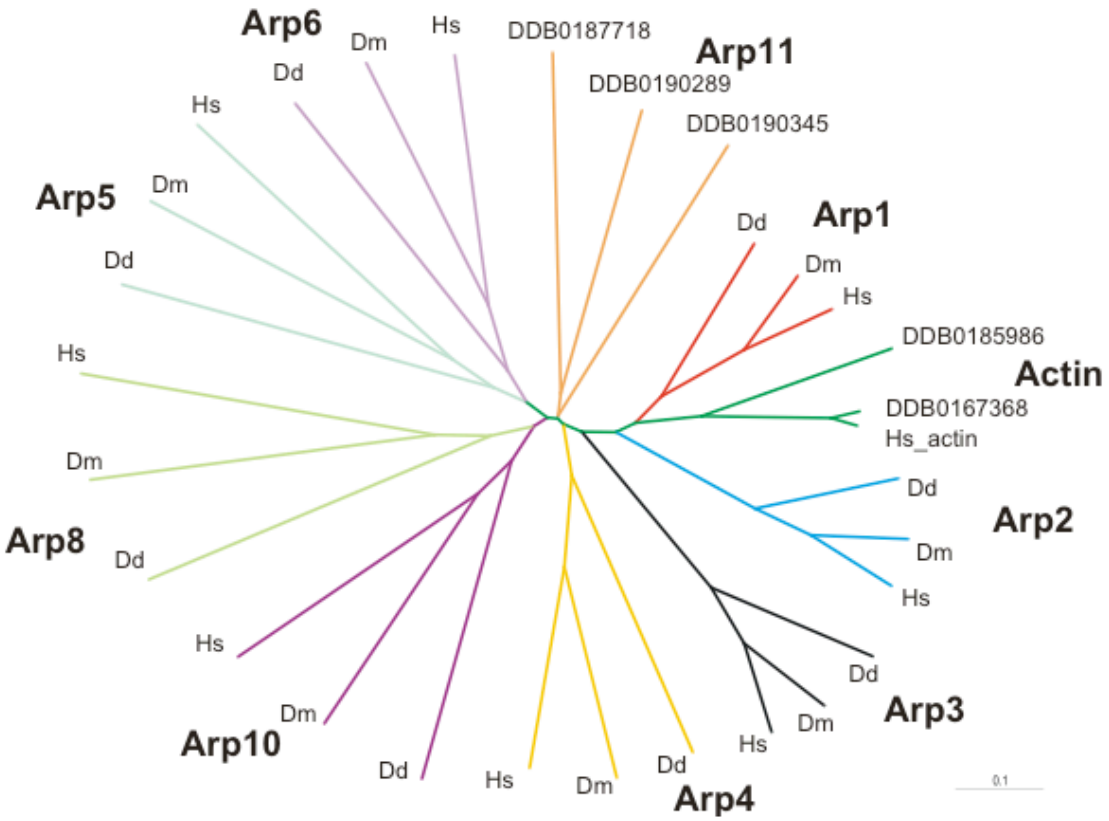


Figure SI 14. Dendrogram of actin and actin-related proteins (Arps).
A CLUSTALX alignment of the Arps from Dictyostelium, Drosophila *and man,* Dictyostelium *filactin (DDB0185986), human ß-actin (Hs_actin) and* Dictyostelium *actin (DDB0167368) was used to create an unrooted dendrogram with the TreeView program. The branches leading to distinct Arp classes have been colour coded. Arp11 proteins appear to be the founding members of a new Arp class, with the numbers corresponding to dictyBase identifiers. Dd:* D. discoideum, *Dm:* D. melanogaster, *Hs:* Homo sapiens. *dictyBase identifiers of the Dictyostelium Arps: Arp1, DDB0188184; Arp2, DDB0168783; Arp3, DDB0218534; Arp4, DDB0215233; Arp5, DDB0184033; Arp6, DDB0187231; Arp8, DDB0187629; Arp10, DDB0187603. The DDB gene identifiers that can be used to locate the annotation for that gene model at dictyBase (http://dictbase.org/) or at geneDB (http://www.genedb.org/genedb/dicty/index.jsp).*

Table SI 13. Proteins involved in Rho signaling and their occurrence in other phyla

Protein Class	Number of genes	Relevant domain or component	Closest relatives in other organisms	Occurence	Notes
Rho GTPases					
Rac-like	6	GTPase	Rac	(P), F, M	1
RhoBTB-like	1	GTPase	RhoBTB	M	2
Other RhoGTPases	8	GTPase	Rac (Unique)	U	3
Dissociation inhibitors					
RhoGDI1	1	RhoGDI	RhoGDI	E	
RhoGDI2	1	RhoGDI	RhoGDI (Unique)	U	4
Exchange factors					
RhoGEF	45	RhoGEF	Mostly unique	F, M	5
DOCK180-related	8	DHR-2	DOCK180, MBC, CED-5	E	6
Darlin	1		SmgGDS, Yeb3p	F, M	
GTPase activating proteins					
RhoGAP	46	RhoGAP	Mostly unique	E	5
Effectors					
Scar complex	5	PIR121	WAVE complex	M	7
WASP	1	PBD	WASP	F, M	
WASP-related	2	PBD	WASP/WAVE (Unique)	U	8
PAK	9	PBD	PAK kinases	F, M	9
Gelsolin-related	1	PBD		U	10
Formins	10	GBD	Formins	(P), F, M	11
IQGAP-related	4	GRD	IQGAP	F, M	12
CIP4	2		CIP4	F, M	13
NADPH oxidase	$\geq$5	$p67^{phox}$	NADPH oxidase	E	14
Class I PI3-kinases	6		Class I PI3-kinases (p110)	(P), (F), M	15
Other lipid kinases	13		PI5K, DGK	E	16
Phospholipase C	1		Phospholipase C	E	
Phospholipase D	10		Phospholipase D	E	

The genome was inspected for domains characteristic of each of the protein classes. Relevant domains refer, apart from the GTPase, to those involved in interactions with the Rho GTPase. For additional effectors that do not display defined domains a list based on recent reviews of the field was elaborated and the *Dictyostelium* genome was interrogated using the BLAST server at dictyBase with the metazoan or fungal protein as query. Rho signaling is an expanding field, therefore no claim of completenes can be made. For many of the protein families included in the table participation in Rho signaling has been documented in some species but not in others, therefore a role in *Dictyostelium* should be taken as putative. Occurrence refers in general to the presence of a protein with equivalent domain architecture. In the cases of large families it just indicates that the relevant domain is represented. **U**, unique (the protein has no relatives in plants, fungi or metazoa, or differs from relatives due to

(*notes continued overleaf)*

an unusual domain composition); **E**, eukaryotes; **P**, plants; **F**, fungi; **M**, metazoa (for E, F, P and M the protein might be missing from any particular species). When in parentheses, occurrence indicates that related proteins with a different domain architecture exist in that particular phylum. Abbreviations: **DHR-2**, Dock homology region 2; **GAP**, GTPase activating protein; **GBD**, Rho GTPase-binding domain of formins; **GDI**, guanine nucleotide dissociation inhibitor; **GEF**, guanine nucleotide exchange factor (the RhoGEF domain is also known as DH, Dbl homology domain); **GRD**, RasGAP-related domain; **PBD**, p21-binding domain (also known as CRIB, Cdc42 and Rac interactive binding).

1. Based on phylogenetic analyses, following Rho GTPases can be grouped in the Rac subfamily: Rac1a, Rac1b, Rac1c, RacF1, RacF2 and RacB.

2. In RhoBTB proteins the GTPase is followed by two BTB domains. In RacA, the *Dictyostelium* RhoBTB orthologue, the GTPase is more closely related to Rac than to the GTPase of metazoan RhoBTB proteins.

3. This group includes RacC to E and RacG to L. RacK is a pseudogene. Named Rac for historical reasons, these proteins have no clear affiliation, although some are closer to Rac than to members of other subfamilies. There are no representatives of the Cdc42 and Rho subfamilies in *Dictyostelium*.

4. RhoGDIs consist of a C-terminal domain with a β-sheet barrel structure and an N-terminal regulatory arm. RhoGDI2 lacks the N-terminal regulatory arm and binding to Rho GTPases has not been demonstrated.

5. The RhoGEFs and RhoGAPs of *Dictyostelium* display a high diversity of domain architectures. Three proteins carry both domains simultaneously. As in other species, the RhoGEF domain is almost invariably followed by a PH domain. Most proteins of these two classes present unique domain combinations and only in very few cases close relatives in metazoa or fungi can be recognized.

6. DOCK180-related proteins function in complex with ELMO1, which in *Dictyostelium* has two putative orthologues.

7. Scar and the members of the complex PIR121, Nap125, Abi2 and HSP300 are each encoded by a single gene.

8. These proteins share a PBD, proline-rich region, WH2 and Arp2/3-binding region with WASP, but lack a WH1 domain.

9. Two of the genes code for an identical protein.

10. Contains two PBDs not found in any other other member of the gelsolin family. They are placed N-terminal to the gelsolin repeats.

11. In general the domain architecture of *Dictyostelium* formins resembles that of the fungal and metazoan orthologues, with an N-terminal Rho-binding domain (GBD) that overlaps with the FH3 domain. Plant formins are devoid of GBD/FH3. One *Dictyostelium* formin, ForI, apparently lacks a recognizable GBD/FH3.

12. The criterium for inclusion in this family is the presence of a RasGAP_C domain, which is present in IQGAPs of other species and is placed immediately downstream of the RasGAP-like domain. The IQGAP-related proteins of *Dictyostelium* lack the N-terminal CH domain characteristic of most metazoan and fungal orthologues. DDB0202463 is an exception: it has two CH domains of the type found in fimbrin.

13. There are two putative CIP4 homologues based on the domain architecture of these proteins, composed of a FCH (Fes/CIP4 homology) domain followed by an SH3 domain.

14. Clear orthologues of the gp91 component (3 genes) and distant relatives of p22 and p67 (each one gene) can be identified in the genome. There might be a weakly conserved unidentified p47 homologue.

15. Class I PI3-kinases of *Dictyostelium* resemble those of metazoa in its domain architecture. One of them lacks a Ras-binding domain and has an N-terminal PH domain.

16. This class includes 8 predicted phosphatidylinositol-4-phosphate 5-kinases and 5 diacylglycerol kinases.

G-protein Coupled Receptors

G-protein coupled cell surface receptors (GPCRs) form the basis of signalling systems in a number of species, allowing the detection of a variety of extracellular signals such as light, Ca^{2+}, odorants, nucleotides and peptides. They are subdivided into six families, but despite their conserved secondary domain structure, they do not share significant sequence similarity across families [43£]. Before genome sequencing began, only the novel cAMP receptor GPCR family had been examined in detail in *Dictyostelium* [44, 45£]. However, given that *Dictyostelium* must respond to numerous signals for feeding and during development, it was not too surprising that a detailed search uncovered 48 additional putative GPCRs (Figure 8, main paper). In addition to twelve CAR/CRL family members, which are distantly related to other GPCRs, there is one potential secretin GPCR (plus one which appears to be a pseudogene), 17 potential metabotropic glutamate/$GABA_B$ receptors and 25 potential frizzled/smoothened GPCRs.

Dictyostelium has three families of GPCRs that have not previously been observed in non-metazoan species. The putative secretin family GPCR is interesting because these proteins were thought to be of relatively recent origin, appearing closer to the time of the divergence of animals [46£]. The *Dictyostelium* protein does not contain the characteristic GPCR proteolytic site, but its transmembrane domain is clearly more closely related to secretin GPCR's than to other families. The *Dictyostelium* family 3 GPCRs are most similar to the mammalian subtype 1 and 2 $GABA_B$ receptors that form heterodimers and regulate Ca^{2+} or K^+ channels through second messengers [47£]. Curiously, some of these $GABA_B$–like receptors appear more closely related to subtype 2 receptors while others display a slightly higher similarity to subtype 1 receptors, suggesting that the *Dictyostelium* receptors may heterodimerise as well. The evidence for the grouping of the putative members of the frizzled/smoothened family is not as strong as for the family 3 GPCRs. Frizzled receptors are defined by a conserved domain structure that consists of an N-terminal cysteine-rich ligand-binding domain (CRD), seven trans-membrane (TM) domains and a short sequence motif (KTXXXW) immediately after the seventh TM domain. The distantly related smoothened receptors lack the KTXXXW motif. For most of the *Dictyostelium* receptors the similarity to the frizzled/smoothened type receptors is restricted to the TM region. However, one of the receptors contains all three critical frizzled domains in the correct arrangement, five others lack only the KTXXXW motif and a further receptor lacks only the CRD. The discovery of a secretin family GPCR, metabotropic $GABA_B$ receptors and distantly related frizzled receptors in *Dictyostelium* suggests that the radiation of GPCR families predates the divergence of the animals and fungi.

SH2 domain proteins

The first *Dictyostelium* SH2 domain-containing proteins to be discovered were three STATs and the genome sequence now reveals there to be one additional STAT, Dd-STATd[48£, 49]. The dictyBase ID numbers for Dd-STATa-d are; DDB0205567, DDB0191116, DDB0192092 and DDB0167449 respectively. *Dictyostelium* also employs the SH2 domain-containing serine-threonine kinase, Shk1, to regulate chemotaxis and the genome sequence reveals four additional SHKs (Shk2-5). The dictyBase ID numbers for Shk1-5 are; DDB0185422, DDB0205433, DDB0188020, DDB0204145 and DDB0219557, respectively. In addition to these known SH2 domain proteins, the whole genome analysis using Pfam predictions identified three other likely SH2-domain containing proteins, giving a total of 12 (Figure SI 15).

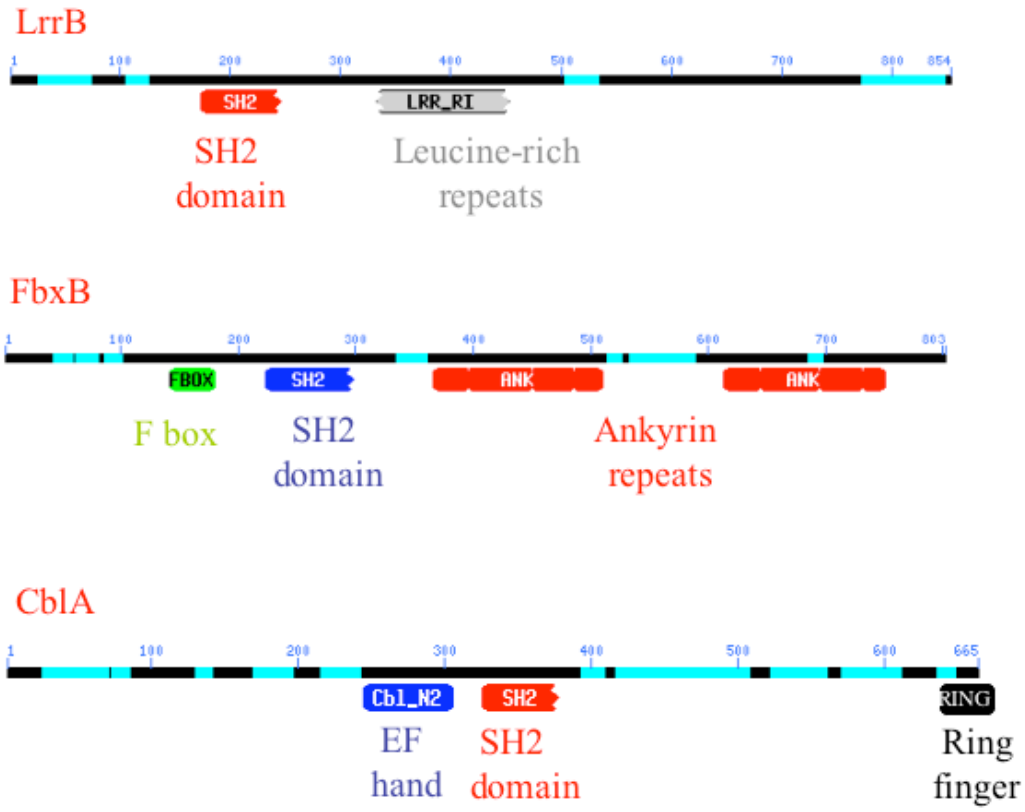


Figure SI 15. New SH2 domain proteins.
The three predicted SH2 domains containing protein sequences were searched against the NCBI protein database and the "conserved domains" are shown. The RING finger of DdCblA, the "conserved domain" was not identified in the conserved domain search but was identified from separate BLAST searches.

One of them CblA is highly related to the metazoan cbl proto-oncogene product (DDB0168162). Cbl is a "RING finger" ubiquitin-protein ligase that recognizes activated receptor tyrosine kinases and various molecular adaptors[15, 19]. The second gene, *fbxB*, encodes an SH2 domain and an "F-box", again a targeting signal for ubiquitinylation (DDB0168622). The third predicted SH2 domain protein, llrB, contains leucine-rich repeats that probably constitute protein-protein interaction domains (DDB0187660). These new SH2 domain proteins broaden the potential scope of SH2 domain signaling in *Dictyostelium.*

Protein Kinases

We classified the protein kinases that we could recognize in *Dictyostelium* based on the Hanks and Hunter classification, as extended by Manning et al.[51£, 52]. The *Dictyostelium* proteome was first scanned for selected PROSITE domains using the ps_scan program from ftp://us.expasy.org/databases/prosite/tools/ps_scan/. Many of the resulting protein sequences contain more than one of the listed domains, resulting in a total of 287 kinase candidates. These proteins were then individually screened for the conserved eukaryotic catalytic domain at http://hodgkin.mbu.iisc.ernet.in/~king/index.html, and analyzed by BLAST at NCBI against the swissprot database. Of the 256 typical eukaryotic protein kinases, 18 do not contain the

catalytic aspartate residue and thus are unlikely to be enzymatically active. Such catalytically inactive kinases have been reported for other organisms, though their function is largely unknown[48£]. However, because the catalytic domains are otherwise conserved, and because these proteins might use a modified catalytic mechanism, we included them in the total count of protein kinases.

Table SI 14. Protein Kinase Domains in *Dictyostelium*

PROSITE Domain	Domain ID	# of domains
PROTEIN_KINASE_DOM	PS50011	256
PROTEIN_KINASE_ATP	PS00107	151
PROTEIN_KINASE_ST	PS00108	181
PROTEIN_KINASE_TYR	PS00109	23
RECEPTOR_TYR_KIN_II	PS00239	2
RECEPTOR_TYR_KIN_III	PS00240	0
RECEPTOR_TYR_KIN_V_1	PS00790	0
RECEPTOR_TYR_KIN_V_2	PS00791	0
MAPK	PS01351	2
HIS_KIN	PS50109	15
PI3_4_KINASE_1	PS00915	9
PI3_4_KINASE_2	PS00916	12
PI3_4_KINASE_3	PS00290	15

The protein kinase-related PROSITE domains that were used to screen the *Dictyostelium* proteome

In addition, the *Dictyostelium* proteome was searched for kinases that do not possess a typical kinase domain (atypical kinases, aPKs). To identify such candidates, human atypical kinase sequences were first BLASTed (BLASTp) against predicted proteins in dictyBase. Sequences with significant hits (E value = e^{-10} over at least 25% of the protein length) were then analyzed further by BLASTp against the Swiss-Prot and non-redundant databases at NCBI, and individual InterPro searches at http://www.ebi.ac.uk/InterProScan/. This resulted in the identification of 23 atypical kinase candidates in addition to the 14 previously identified histidine kinases that also contain an atypical protein kinase domain.

Seventy-three of the protein kinases have been previously characterized, including well-conserved members of the p21-activated kinase (PAK) family, the cyclin-dependent kinase (CDK) family, the PKAs (cAMP-dependent protein kinases), the MAP kinases ErkA and ErkB, the glycogen synthase kinase III (GSK3) GskA, myosin light chain kinase, and casein kinases I and II (CKI, CKII), as well as the atypical histidine kinases and the myosin heavy chain kinases.

Table SI 15. Protein kinases of *Dictyostelium*

Group	Dicty	Yeast*	Worm*	Fly*	Human*
AGC	23	17	30	30	63
CAMK	24	21	46	32	74
CK1	2	4	85	10	12
CMGC	33	21	49	33	61
HisKin	14	1	0	0	0
Other	62	38	67	45	83
STE	42	14	25	18	47
Tyrosine kinase	0	0	90	32	90
Tyrosine kinase-like	72	0	15	17	43
RGC	0	0	27	6	5
Atypical-A6	1	1	2	1	2
Atypical-ABC1	4	3	3	3	5
Atypical-Alpha	6	0	4	1	6
Atypical-BRD	1	0	1	1	4
Atypical-G11	1	0	0	0	1
Atypical-PIKK	5	5	5	5	6
Atypical-RIO	2	2	3	3	3
Atypical-TAF1	1	1	1	1	2
Atypical-TIF	2	0	0	0	3
Atypical-Other	0	2	1	1	7
Total	295	125	449	235	512

AGC: PKA, PKG, and PKC families
CAMK: Ca2+/CAM-dependent protein kinases
CK1: Casein kinase I
CMCG: CDK, MAPK, GSK3, CLK families
HisKin: (two-component) Histidine Kinases
Other: Kinases that do not fit into any other group; CK2 (casein kinase II), Nek, PEK families
STE: Homologs of yeast sterile 7, 11, 20 kinases, contain PAK (p20-activated) kinases
TK: Tyrosine kinase
TKL: Tyrosine kinase-like; MLK (mixed lineage kinase), RAF, LIMK families
RGC: Receptor guanylate cyclase
Atypical Kinases: No homology to conventional protein kinases
Atypical-A6: Twinfilin/PTK9 family
Atypical-ABC1: Activity of BC(1) complex; required for coenzyme Q (ubiquinone) biosynthesis
Atypical Alpha: *Dictyostelium* MHCKs; eukaryotic elongation factor 2 kinases
Atypical-BRD: Bromodomain protein
Atypical-G11: Uncharacterized
Atypical-RIO: *Ri*ght *O*pen reading frame; processing of 20S pre-rRNA
Atypical-TAF1: TATA box binding protein (TBP) associated factor
Atypical-TIF: Transcriptional Intermediary Factor 1 (human)
Atypical-Other: FAST, H11, PDHK families
* The protein kinase complement of the human genome. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) *Science* 298 (5600); 1912-34.

A total of 24 putative Ca^{2+}/calmodulin-dependent kinases (CAMK) were identified. We found four putative protein kinases that are very similar to the known myosin light chain kinase mlkA, bringing the number of CAMK/MLCK-classified kinases in *Dictyostelium* to five. The same number was found for worm, fly, and human (Table SI 15). As reported by others recently, we also identified three additional putative myosin heavy chain kinases[40]. These belong to the atypical Alpha kinase family and are similar to the cloned heavy chain kinases MhkA, MhkB, and MhkC. Members of other atypical kinase families share high sequence similarity with human aPKs, in particular, the A6, ABC1, and RIO kinase families (Table SI 16). In worm, human, and *S. pombe*, three members of the wee1 family of cell cycle kinases exist: Wee1, Mik1, and Myt1. Interestingly, we identified three wee1-related kinases in *Dictyostelium*. There is also a putative second casein kinase II with high similarity to *Dictyostelium* casK and other CKIIs, a second CKI, similar to *Dictyostelium* Cak1 and a second GSK-3. In addition to the four that have been previously characterized, we identified eight additional CDKs. The 12 putative CDKs are comparable to the number found in animals.

A complete list of *Dictyostelium* protein kinases can be found at http://dictybase.org/GeneFamilies/ProteinKinases.html.

Table SI 16. Newly identified protein kinases similar to proteins in other species.

dictyBase ID	Group / Family / Subfamily	Top BLASTP hit (species)	% identity / % protein length
DDB0220670	AGC / AKT	AKT2 (H. sapiens)	43 / 55
DDB0219947	AGC / NDR	CBK1 (S. cerevisiae)	43 / 99
DDB0216241	AGC / NDR	CBK1 (S. cerevisiae)	44 / 94
DDB0216243	AGC / PDK	PDPK1 (R. norvegicus)	32 / 98
DDB0216238	CAMK / CAMKL / MARK	MARK4 (H. sapiens)	32 / 70
DDB0216369	CAMK / CAMKL / MARK	MARK4 (H. sapiens)	26 / 93
DDB0216307	CAMK / MLCK	CAMK1 (R. norvegicus)	47 / 85
DDB0216312	CAMK / MLCK	CAMK1 (H. sapiens)	45 / 95
DDB0216309	CAMK / MLCK	CAMK1 (M. musculus)	48 / 84
DDB0216308	CAMK / MLCK	CAMK1 (H. sapiens)	47 / 76
DDB0216333	CAMK / RAD53	CHK2 (M. musculus)	35 / 66
DDB0216336	CK1 / TTBK	KC1E (H. sapiens)	37 / 65
DDB0229427	CMGC / CDK	CDK10 (H. sapiens)	52 / 84
DDB0216376	CMGC / CDK	CDK10 (H. sapiens)	51 / 85
DDB0229428	CMGC / CDK / CRK7	CRK7 (H. sapiens)	46 / 58
DDB0229429	CMGC / DYRK / DYRK1	DYRA (R. norvegicus)	45 / 68
DDB0216280	CMGC / GSK	GSK3B (H. sapiens)	35 / 81
DDB0216281	CMGC / PRP4	PR4B (H. sapiens)	35 / 97
DDB0229430	CMGC / RCK	MAK (M. musculus)	59 / 54
DDB0216254	OTHER / AUR	STK6 (M. musculus)	48 / 92
DDB0219953	OTHER / CK2	CSNK2A1 (R. norvegicus)	57 / 56
DDB0216282	OTHER / NEK	Nek2 (H. sapiens)	48 / 80
DDB0229408	STE / STE20	STK3 (H. sapiens)	40 / 72
DDB0216377	STE / STE20	STK3 (H. sapiens)	44 / 57
DDB0216374	STE / STE20	STK3 (H. sapiens)	40 / 78
DDB0167076	STE / STE20 / YSK	CC7 (S. pombe)	33 / 98
DDB0216426	Atypical / A6	A6 (H. sapiens)	33 / 100
DDB0216427	Atypical / ABC1	ADCK4 (H. sapiens)	38 / 83
DDB0216431	Atypical / ABC1	ADCK5 (H. sapiens)	41 / 64
DDB0216428	Atypical / RIO	RIOK1 (H. sapiens)	50 / 80
DDB0216429	Atypical / RIO	RIOK2 (H. sapiens)	50 / 84
DDB0229332	Atypical / PIKK / ATR	ATR (H. sapiens)	28 / 65
DDB0229336	Atypical / PIKK / DNAPK	DNPK1 (H. sapiens)	28 / 100
DDB0229297	Atypical / PIKK / SMG1	SMG1 (H. sapiens)	25 / 81
DDB0229338	Atypical / PIKK / TRRAP	TRRAP SMG1 (H. sapiens)	22 / 100

Dictyostelium gene predictions in which the ratio of % identity to % protein length is either ≥20 / ≥65 or ≥40 / ≥50. Classification is based on comparison of the entire length of the proteins.

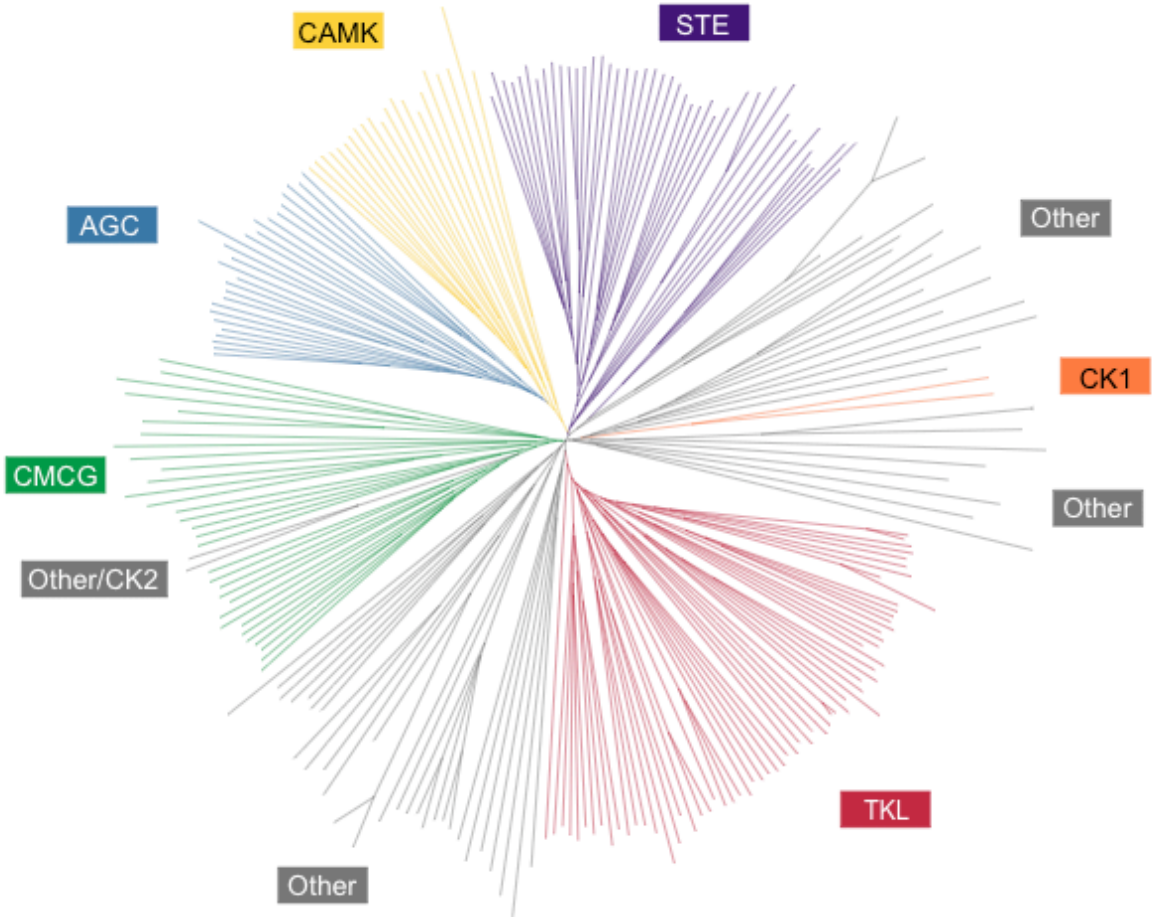


Figure S1 16. *Dictyostelium* protein kinase dendrogram.
A CLUSTALX alignment of the catalytic domains of all putative protein kinases in the Dictyostelium genome was used to create a dendrogram in the TreeView program.

Transcription factors

Initial descriptions of the transcriptional profile of developing *Dictyostelium* cells have been published[11, 15, 49£, 50]. Genes expressed in specific cell types have been identified and many of these have been confirmed by in situ hybridization experiments that reveal an unexpected level of complexity in the spatial control of gene expression[11]. The genome sequence now provides the means to explore the cis-acting regions that control this transcriptional activity and provides the set of proteins required for its regulation. One hundred and six proteins contain protein domains common to transcription factors and 93 of these appear to be authentic transcription factors (Table SI 17; Figure SI 17). This is close to the 149 found in *S. cerevisiae*. The most common domain is the Myb DNA binding domain, followed by the GATA type Zn fingers, basic-leucine zipper (bZIP), homeobox, p53-like and MADS-box domains. There is also one member each of the histone-like transcription factor CBF/NF-Y/archaeal histone subunit A, histone-like transcription factor CBF/NF-Y/archaeal histone subunit B, and DNA-binding WRKY. Interestingly, no proteins having a basic helix-loop-helix could be recognized although this motif is found in all kingdoms. Most of the previously characterized *Dictyostelium* transcription factors were identified in the genome sequence: hbx2, WarA

(homeobox-containing proteins), MybA, MybB, MybC (myb-domain), STATa, STATb, STATc, STATd (p53-like DNA binding domain), ComH, StkA (GATA-Zn finger), DimA (basic leucine zipper), and SrfA (MADS box).

Dictyostelium appears to have fewer transcription factors than are found in the metazoa (Table SI 18; Figure SI 18). However, since we identified the transcription factors by known DNA binding domains, the analysis is heavily weighted towards the metazoa and those factors more specific to the amoebozoa would have been missed. Supporting this idea, some known *Dictyostelium* transcription factors were not identified in the current analysis; CRTF, CbfA and GBF[12, 13, £53]. CbfA binds DNA through an AT hook motif, which is usually found in chromatin remodeling factors. Biochemical evidence suggests that CRTF and GBF bind DNA through atypical zinc finger motifs.

Methodology. The *Dictyostelium* proteome was scanned for PROSITE domains using the ps_scan program from ftp://us.expasy.org/databases/prosite/tools/ps_scan/. The number of proteins containing these domains in other species was obtained from EBI (example URL: http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-id+ucin1NEKC9+-e+[INTERPRO:IPR001092])

Most of the proteins having transcription factor domains had some sequence similarity with protein sequences from the SwissProt database. The complete list of transcription factors can be found at: http://dictybase.org/GeneFamilies/TranscriptionMachinery.html.

Table SI 17. Predicted *Dictyostelium* proteins containing transcription factor domains

Domain Name	Domain ID	Found in (# of proteins)	Total number of occurrences of the domain
BZIP	IPR004827	19	19
CBFA_NFYB	IPR003956	1	1
CBFB_NFYA	IPR001289	5	5
E2F_TDP	IPR003316	2	2
Fungi_Trscrp_N	IPR001138	2	2
GATA_ZN_FINGER	IPR000679	19	19
HLH	IPR001092	0	0
HOMEOBOX	IPR001356	13	15
HSF_DNA-bind	IPR000232	1	1
HTH	IPR001387	1	1
MADS	IPR002100	4	4
MYB	IPR001005	28	47
NmrA	IPR008030	1	1
p53-like	IPR008967	4	4
PAH	IPR003822	1	1
PC4	IPR003173	2	2
SART-1	IPR005011	1	1
WRKY	IPR003657	1	2
zf-NF-X1	IPR000967	1	1
TOTAL:		**106**	**128**

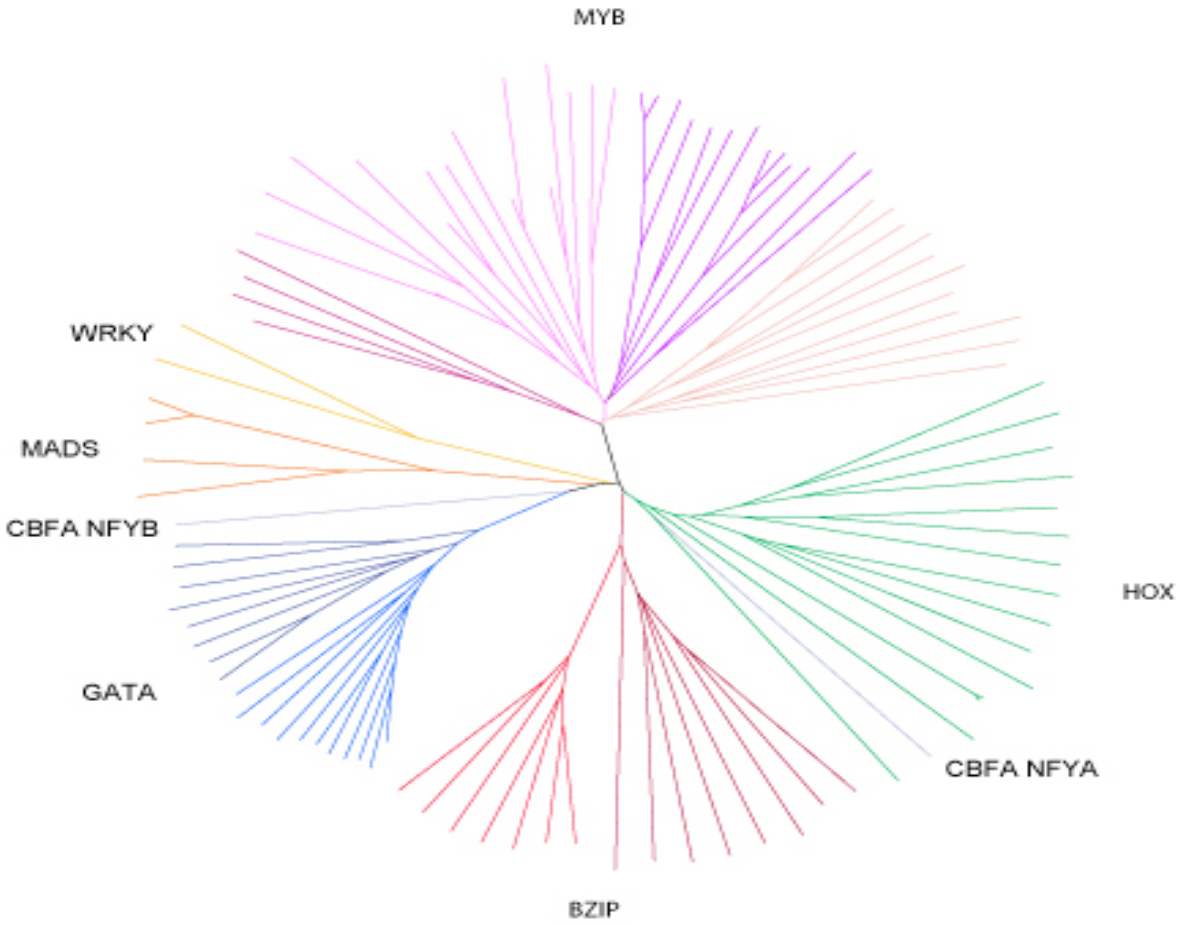


Figure SI 17. *Dictyostelium* transcription factor dendrogram.
A CLUSTALX alignment of the domains of all putative transcription factors in the *Dictyostelium* genome was used to create a dendrogram in the TreeView program.

Table SI 18. Transcription factor domains in *Dictyostelium* and other species

Domain Name	Domain ID	Dd[1]	Sc	Dm	Ce	Hs	Mm	At
BZIP	IPR004827	19	18	48	36	98	93	135
CBFA_NFYB	IPR003956	1	1	0	0	1	2	9
CBFB_NFYA	IPR001289	5	1	1	2	2	3	18
E2F_TDP	IPR003316	2	0	7	5	11	24	22
GATA	IPR000679	19	11	18	18	23	23	41
HLH	IPR001092	0	10	94	59	181	196	269
HOMEOBOX	IPR001356	13	10	187	126	325	386	147
Hrmon_recept_lig	IPR000536	0	0	43	358	81	80	0
HSF_DNA-bind	IPR000232	1	5	2	2	7	9	31
HTH	IPR001387	1	1	1	1	3	2	4
MADS	IPR002100	4	4	5	3	9	16	165
MYB	IPR001005	28	19	50	25	92	76	602
NmrA	IPR008030	1	0	0	0	0	0	0
p53-like	IPR008967	4	3	37	36	91	124	5
PAH	IPR003822	1	1	8	1	5	9	35
PC4	IPR003173	2	1	2	1	1	1	3
SART-1	IPR005011	1	1	3	1	1	4	3
WRKY	IPR003657	1	0	0	0	0	0	75
Zn_clus	IPR001138	2	62	0	0	0	0	0
zf-NF-X1	IPR000967	1	1	2	2	8	5	2
TOTAL:		**106**	**149**	**508**	**676**	**939**	**1053**	**1566**

[1]Abbreviations: Dd: *Dictyostelium discoideum*, Sc: *Saccharomyces cerevisiae*, Dm: *Drosophila melanogaster*, Ce: *Caenorhabditis elegans*, Hs: *Homo sapiens*, Mm: *Mus musculus*, At: *Arabidopsis thaliana*.

BZIP: basic-leucine zipper
CBFA_NFYB: CCAAT-binding transcription factor, subunit B
CBFB_NFYA: Histone-like transcription factor CBF/NF-Y/archaeal histone, subunit A
E2F_TDP: Transcription factor E2F/dimerisation partner (TDP)
GATA: Zn-finger, GATA type
Hrmon_recept_lig: Ligand-binding region of nuclear hormone receptor
HLH: Basic helix-loop-helix dimerisation region bHLH
HOX: Homeobox protein
HSF_DNA-bind: Heat shock factor (HSF)-type, DNA-binding
HTH: Helix-turn-helix motif
MADS: Transcription factor, MADS-box/serum response factor (SRF)
MYB: Myb, DNA-binding
NmrA: nitrogen metabolite repression
p53-like: p53-like transcription factor, DNA-binding
PAH: Paired amphipathic helix
PC4: Transcriptional coactivator p15
SART-1: Leucine zipper protein
WRKY: defined by the conserved amino acid sequence WRKYGQK at its N-terminal end
Zn_clus: Fungal transcriptional regulatory protein, N-terminal
zf-NF-X1: Zn-finger, NF-X1 type

Pfam domain analysis of transcription factors

Using the strategy described above for determining the presence and absence of Pfam domains in eukaryotes, we examined the domains found in the transcription factors. Of the 159 transcription factor families, 100 are eukaryote-specific. The remaining 59 are predominantly bacterial or archeal and most of them are not present in eukaryotes.
Of the 100 eukaryote-specific transcription factor Pfam domains:
- 23 are present in all four clades (plant, fungi, animal, amoebozoa)
- 39 are animal-specific
- 12 are plant-specific
- 9 are fungi-specific
- 0 are *Dictyostelium*-specific (the *Dictyostelium* proteins have not yet been designated as distinct Pfam models)
- 17 are present in two or three clades

The 28 common transcription factors (including 5 not present in *Dictyostelium*) make up the following fractions of the total transcription factors that can be identified in each organism:

Dictyostelium	77%
Yeasts	52%
Aspergillus	22%
Neurospora	36%
Plants	57%
C. elegans	34%
Human	34%

The clade-specific families make up the following proportion of the transcription factor proteins in each genome:

Yeasts	20%
Aspergillus	38%
Neurospora	27%
Plants	30%
C. elegans	55%
Human	20%

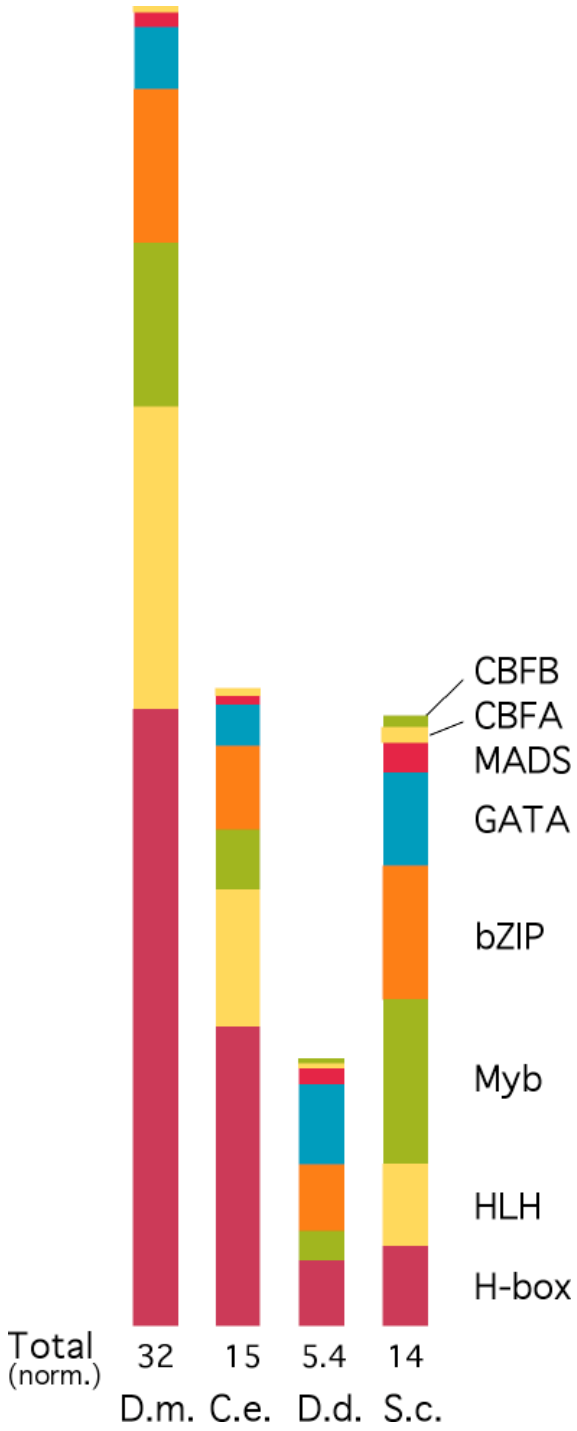


Figure SI 18. Relative occurrence of transcription factor families.
The number of proteins that contain major transcription factor domains are represented by the length of the differently colored bars. The numbers were normalized to the total number of genes in each organism (D.m., Drosophila *13,473; C.e.,* C. elegans *19,173; D.d,* Dictyostelium *12,500; S.c.,* S. cerevisiae *5,538). Total gene numbers were obtained from: http://flybase.bio.indiana.edu/ for* Drosophila*, http://www.wormgenes.org/ for the worm, and reference[54£] for* S. cerevisiae*.*

Methods

Sequencing and assembly

Generally, chromosome-specific HAPPY markers, mapped genes, or reads derived from YACs whose locations were confirmed by the HAPPY map, were used as seeds to nucleate bins of reads from the impure WCS libraries into chromosome specific subsets via BLAST[55£] or Atlas-overlapper[56£]. These subsets were then assembled into contigs using either GAP4[57£] (chromosomes 1-3) or the PHRED/PHRAP/CONSED package (chromosomes 4-6; http://phrap.org). Read-pair information, BLAST searches and reassembly were then used iteratively to identify further chromosome-specific reads from all available sequence data to extend the contigs and link them into scaffolds (groups of two or more contigs linked by robust read-pair information). The chromosomal origin of the resulting contigs was verified based on their proportional content of reads from the respective libraries (KS, unpublished software). The main differences in strategy between chromosomes 1-3 and 4-6 were that, for the former, initial contig seeding was mainly from chromosomally assigned gene sequences and the HAPPY map was used extensively to guide assembly. For the latter, chromosomally assigned HAPPY markers were the main contig seeds whilst YAC-derived reads guided regional assembly; the HAPPY map was reserved until assembly was complete to provide independent verification. Southern mapping was also used to validate assembly in some instances. In C2 one gap remained which was spanned only by the genetic map. This information was used to orientate the two segments of the chromosome.

Gap-closing strategies included primer-walking on available pUC clones, with sub-cloning[58£] and transposon insertion[59£] being used to sequence difficult (generally A+T-rich) templates. PCR products were generated as sequencing templates across gaps which were defined precisely by map data. Complex repetitive regions were resolved by inspecting minor polymorphisms in the repeat sequence, by long-range PCR between flanking non-repetitive sequences, or by the use of YAC-derived (hence region-specific) reads.

A 1.5Mbp portion of chromosome 6 ('EUDICT region') was sequenced as a pilot project using a variety of approaches, sharing the HAPPY map as a common framework. The majority of this region was assembled using the packages described above (but also using the Phusion assembler[60£] supplemented by read-pair processing by Cyclops - http://intweb.sanger.ac.uk/Software/sequencing/docs/harper/cyclops.shtml), whilst two segments were assembled by shotgun sequencing of YAC clones after validation of their location and integrity against an earlier HAPPY map[61£]. There was one major difference in approach in that no bins were created using blast but that all reads from each centre were pooled and assembled with phrap. The seed contigs for the EUDICT region were selected from the resulting assembly through presence of HAPPY map markers or reads from YACs confirmed by evidence from the HAPPY map to be from the region of interest.

Gene prediction and identification of sequence features

Gene prediction was performed using GeneID[62£], HMMGene[63£] and Genefinder (P. Green, unpublished), each trained on a similar but not identical set of well-characterized *D. discoideum* genes. Predictions of all three packages were integrated using GFMerge (S. Spiegler, unpublished) which derives a consensus set of predictions based on concordance amongst the predictions and their concordance with other available data including similarity to *D. discoideum* cDNA sequences[64£] and homology to UniProt entries. cDNA similarities were

identified using exonerate (G. Slater, in preparation) after masking repetitive and low-complexity regions using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and Dust (Tatusov, R. L. and D. J. Lipman, in preparation); comparisons were performed both against the individual cDNA sequences and against the *D. discoideum* UniGene clusters using a customized parameter set (G. Slater, personal communication). Homologies to UniProt entries were found using Washington University BLAST 2.0 blastx (WU-BLAST 2.0, W. Gish unpublished) with softmasking and with a customized parameter set (W=6 WINK=6 nogap), with cut-offs of $e<10^{-10}$, score >200 and identity >20%. Signal peptides and transmembrane helices were predicted on the translated gene models selected by GFMerge, using SignalP and TMHMM[65£, 66£]. GPI-anchor predictions on the translated gene models selected by GFMerge were done using DGPI developed by Julien Kronegg and Didier Buloz (retrieved 31-03-2004 from http://129.194.185.165/dgpi/). Iprscan was used to compare the translated gene models selected by GFMerge against InterPro (release 7) with the default settings and the interpro2go setting so that Gene Ontology terms were automatically assigned[67£]. The GOtcha search was carried out separately to supplement the Gene Ontology assignments by interpro2go (D. Martin, unpublished). All gene product models that are explicitly described were examined manually for accuracy. The Pfam and Interpro lists at http://www.genedb.org/genedb/dicty/index.jsp were inspected for domains characteristic of each of the protein classes. The *Dictyostelium* proteome was also scanned with selected PROSITE domains using the ps_scan program from ftp://us.expasy.org/databases/prosite/tools/ps_scan/. Members from each of these groups were selected to search the genome by BLAST for proteins for which the automatic annotation had failed to recognise the corresponding domains. New candidate proteins were examined manually and against the Pfam and Superfamily gene models for validity. For those proteins that did not display defined domains the *Dictyostelium* genome was interrogated using the BLAST server at dictyBase. For this, amino acid sequences of known *Dictyostelium* proteins were used to search for previously unidentified homologues. Amino acid sequences of additional vertebrate, fungal and plant actin-binding proteins and Arps were used to search for homologues in *Dictyostelium*. Occurrence of a particular *Dictyostelium* actin-binding protein in other phyla was investigated using the architecture analysis tool at the SMART server (http://smart.embl-heidelberg.de/) or BLAST at NCBI against specific databases. The domain composition of each protein was analysed using the sequence analysis tool at the SMART server or the domain distributions found on geneDB.

Blastn (NCBI-blast 2.2.8 with parameters -W 30 -G 2 -F "m D" and minimum score cut-off) was used to find the locations of known *D. discoideum* repeat element sequences, HAPPY markers and rDNA palindrome-related features in the genome sequence. tRNAs were detected using tRNAscan-SE[6].

Availability of reagents

A set of plasmid clones representing a minimum tiling path covering >90% of the genome, as well as the reference strain of *D. discoideum* used in this project (Ax4-1986) will be made available via the Dicty Stock Centre (http://dictybase.org/StockCenter/StockCenter.html). In the interim, requests for reagents should be directed to A.K. (akuspa@bcm.tmc.edu) or A.A.N. (Noegel@Uni-Koeln.DE).

References

1. Kohrle, J. Brigelius-Flohe, R., Bock, A., Gartner, R., Meyer, O., Flohe L. Selenium in biology: facts and medical perspectives. *Biol. Chem.* **381**, 849-864 (2000).
2. Bosl, M. R., Takaku, K., Oshima, M., Nishimura, S. & Taketo, M. M. Early embryonic lethality caused by targeted disruption of the mouse selenocysteine tRNA gene (Trsp). *Proc Natl Acad Sci U S A* **94**, 5531-4 (1997).
3. Fu, L. H. et al. A selenoprotein in the plant kingdom. Mass spectrometry confirms that an opal codon (UGA) encodes selenocysteine in Chlamydomonas reinhardtii gluththione peroxidase. *J Biol Chem* **277**, 25983-91 (2002).
4. Novoselov, S. V. et al. Selenoproteins and selenocysteine insertion system in the model plant cell system, Chlamydomonas reinhardtii. *Embo J* **21**, 3681-93 (2002).
5. Osaka, T. et al. The protozoa dinoflagellate Oxyrrhis marina contains selenoproteins and the relevant translation apparatus. *Biochem Biophys Res Commun* **300**, 236-40 (2003).
6. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).
7. Guimaraes, M. J. et al. Identification of a novel selD homolog from eukaryotes, bacteria, and archaea: is there an autoregulatory mechanism in selenocysteine metabolism? *Proc Natl Acad Sci U S A* **93**, 15086-91 (1996).
8. Krol, A. Evolutionarily different RNA motifs and RNA-protein complexes to achieve selenoprotein synthesis. *Biochimie* **84**, 765-74 (2002).
9. Lambert, A., Lescure, A. & Gautheret, D. A survey of metazoan selenocysteine insertion sequences. *Biochimie* **84**, 953-9 (2002).
10. Kryukov, G. V. et al. Characterization of mammalian selenoproteomes. *Science* **300**, 1439-43 (2003).
11. Maeda, M. et al. Changing patterns of gene expression in *Dictyostelium* prestalk cell subtypes recognized by in situ hybridization with genes from microarray analyses. *Eukaryot Cell* **2**, 627-37 (2003).
12. Schnitzler, G. R., Fischer, W. H. & Firtel, R. A. Cloning and characterization of the G-box binding factor, an essential component of the developmental switch between early and late development in *Dictyostelium*. *Genes Devel.* **8**, 502-514 (1994).
13. Mu, X. Q., Spanos, S. A., Shiloach, J. & Kimmel, A. CRTF is a novel transcription factor that regulates multiple stages of *Dictyostelium* development. *Development* **128**, 2569-2579 (2001).
14. Kimmel, A. R. & Firtel, R. A. Sequence organization and developmental expression of an interspersed, repetitive element and associated single-copy DNA sequences in

Dictyostelium discoideum. *Mol Cell Biol* **5**, 2123-30 (1985).
15. VanDriessche, N. et al. A transcriptional profile of multicellular development in *Dictyostelium discoideum*. *Development* **129**, 1543-1552 (2002).
16. Olsen, R. & Loomis, W. F. A model of orthologous protein sequence divergence. *Journal of Molecular Evolution* in press (2004).
17. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-40 (1995).
18. Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. & Gough, J. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res* **32** Database issue, D235-9 (2004).
19. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-20 (1997).
20. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res* **32** Database issue, D138-41 (2004).
21. Thomason, P. & Kay, R. Eukaryotic signal transduction via histidine-aspartate phosphorelay. *J. Cell Sci.* **113**, 3141-3150 (2000).
22. Kim, L., Liu, J. C. & Kimmel, A. R. The novel tyrosine kinase ZAK1 activates GSK3 to direct cell fate specification. *Cell* **99**, 399-408 (1999).
23. Roelofs, J. & Van Haastert, P. J. Genes lost during evolution. *Nature* **411**, 1013-4 (2001).
24. Salzberg, S. L., White, O., Peterson, J. & Eisen, J. A. Microbial genes in the human genome: lateral transfer or gene loss? *Science* **292**, 1903-6 (2001).
25. Stanhope, M. J. et al. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411**, 940-4 (2001).
26. Deivanayagam, C. C. et al. Novel fold and assembly of the repetitive B region of the Staphylococcus aureus collagen-binding surface protein. *Structure Fold Des* **8**, 67-78 (2000).
27. Stechmann, A. & Cavalier-Smith, T. The root of the eukaryote tree pinpointed. *Curr Biol* **13**, R665-6 (2003).
28. Freeze, H. & Loomis, W. F. Chemical analysis of stalk components of *Dictyostelium discoideum*. *Biochim. Biophys. Acta* **539**, 529-537 (1978).
29. Zhang, P., McGlynn, A., Loomis, W. F., Blanton, R. L. & West, C. M. Spore coat formation and timely sporulation depend on cellulose in *Dictyostelium*. *Differentiation* **67**, 72-79 (2001).
30. Metcalf, T., Kelley, K., Erdos, G. W., Kaplan, L. & West, C. M. Formation of the outer layer of the *Dictyostelium* spore coat depends on the inner-layer protein SP85/PsB. *Microbiology* **149**, 305-317 (2003).
31. Blanton, R. L., Fuller, D., Iranfar, N., Grimson, M. J. & Loomis, W. F. The cellulose synthase gene of *Dictyostelium*. *Proc Natl Acad Sci U S A* 97, 2391-6 (2000).
32. Dehal, P. et al. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157-67 (2002).
33. Ennis, H. L. et al. in *Dictyostelium - A model system for cell and developmental biology*. (eds. Maeda, Y., Inouye, K. & Takeuchi, I.) 393-407 (Universal Academy Press, Tokyo, Japan, 1997).
34. Li, Y. et al. Plant expansins are a complex multigene family with an ancient evolutionary origin. *Plant Physiol.* **128**, 854-864 (2002).

35. Robinson, V. & Williams, J. A marker of terminal stalk cell terminal differentiation in *Dictyostelium*. *Differentiation* **61**, 223-228 (1997).
36. Wang, Y. Z., Slade, M. B., Gooley, A. A., Atwell, B. J. & Williams, K. L. Cellulose-binding modules from extracellular matrix proteins of *Dictyostelium discoideum* stalk and sheath. *Eur. J. Biochem.* **268**, 4334-4345 (2001).
37. Witke, W., Schleicher, M. & Noegel, A. A. Redundancy in the microfilament system - abnormal development of *Dictyostelium* cells lacking two F-actin cross-linking proteins. *Cell* **68**, 53-62 (1992).
38. McKeown, M. & Firtel, R. A. Differential expression and 5'end mapping of actin genes in *Dictyostelium*. *Cell* **24**, 799-807 (1981).
39. Romans, P., Firtel, R. A. & Saxe III, C. L. Gene-specific expression on the actin multigene family of *Dictyostelium discoideum*. *J. Mol. Biol.* **186**, 337-355 (1985).
40. Tsang, A. S., Mahbubani, H. & Williams, J. G. Cell-type-specific actin mRNA populations in *Dictyostelium discoideum*. *Cell* **31**, 375-382 (1982).
41. Knecht, D. A., Cohen, S. M., Loomis, W. F. & Lodish, H. F. Developmental regulation of *Dictyostelium discoideum* actin gene fusions carried on low-copy and high-copy transformation vectors. *Mol. Cell. Biol.* **6**, 3973-3983 (1986).
42. Merlot, S. and Firtel, R.A. Leading the way: directional sensing through phosphatidylinositol 3-kinase and other signalling pathways. *J. Cell Sci.* **116**, 3471-3478 (2003)
43. Bockaert, J. & Pin, J. P. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *Embo J* **18**, 1723-9 (1999).
44. Ginsburg, G. T. et al. The regulation of *Dictyostelium* development by transmembrane signalling. *J Eukaryot Microbiol* **42**, 200-5 (1995).
45. Raisley, B., Zhang, M., Hereld, D. & Hadwiger, J. A. A cAMP receptor-like G protein-coupled receptor with roles in growth regulation and development. *Dev Biol* **265**, 433-45 (2004).
46. King, N., Hittinger, C. T. & Carroll, S. B. Evolution of key cell signaling and adhesion protein families predates animal origins. *Science* 301, 361-3 (2003).
47. Hammond, D. L. GABA(B) receptors: new tricks by an old dog. *Curr Opin Pharmacol* **1**, 26-30 (2001).
48. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912-34 (2002).
49. Iranfar, N. et al. Expression patterns of cell-type-specific genes in *Dictyostelium*. *Mol Biol Cell* **12**, 2590-600 (2001).
50. Iranfar, N., Fuller, D. & Loomis, W. F. Genome-wide expression analyses of gene regulation during early development of *Dictyostelium discoideum*. *Eukaryot Cell* **2**, 664-70 (2003).
51. Hanks, S. K. & Hunter, T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *Faseb J* **9**, 576-96 (1995).
52. Manning, G., Plowman, G. D., Hunter, T. & Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* **27**, 514-20 (2002).
53. Winckler, T. et al. CbfA, the C-module DNA-binding factor, plays an essential role in the initiation of development of *Dictyostelium*. *Eukaryotic Cell* in press (2004).
54. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**,

241-54 (2003).
55. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).
56. Havlak, P. et al. The Atlas genome assembly system. *Genome Res*. **14,** 721-32 (2004).
57. Staden, R., Beal, K. F. & Bonfield, J. K. The Staden package, 1998. *Methods Mol. Biol.* **132,** 115-130 (2000).
58. Quail, M. A. M13 cloning of mung bean nuclease digested PCR fragments as a means of gap closure within A/T-rich, genome sequencing projects. *DNA Seq.* **12,** 355-359 (2001).
59. Devine, S. E., Chissoe, S. L., Eby, Y., Wilson, R. K. & Boeke, J. D. A transposon-based strategy for sequencing repetitive DNA in eukaryotic genomes. *Genome Res.* **7,** 551-63 (1997).
60. Mullikin, J. C. & Ning, Z. The phusion assembler. *Genome Res.* **13,** 81-90 (2003)
61. Konfortov, B. A., Cohen, H. M., Bankier, A. T. & Dear, P. H. A high-resolution HAPPY map of *Dictyostelium discoideum* chromosome 6. *Genome Res*. **10,** 1737-1742 (2000).
62. Parra, G., Blanco, E. & Guigo, R. GeneID in *Drosophila*. *Genome Res.* **10,** 511–515 (2000).
63. Krogh, A. Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **5,** 179-186 (1997).
64. Urushihara, H. et al. Analyses of cDNAs from growth and slug stages of *Dictyostelium discoideum*. *Nucleic Acids Res*. **32,** 1647-53 (2004).
65. Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**, 1-6 (1997).
66. Nielsen, H. & Krogh, A. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* **6**, 122-30 (1998).
67. Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-8 (2001).